TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Bioinformatik

Next Generation Machine Learning Prediction of Protein Cellular Sorting

Tatyana Goldberg

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. Daniel Cremers

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Burkhard Rost
2. Prof. Dr. Yana Bromberg, The State University of New Jersey/USA
3. Univ.-Prof. Dr. Iris Antes

Die Dissertation wurde am 15.12.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 22.03.2016 angenommen.

Table of Contents

Abstract

Living cells are divided into specific compartments, each responsible for a different cellular function. The identification of protein localization (*i.e.* into which sub-cellular compartment it is being sorted) is important for understanding protein function, as certain functions can only be performed in certain environments. Despite advances in high-throughput experiments for protein localization, the gap between the number of known proteins and the number of proteins with known localization continues to grow. Several computational approaches have been developed to predict protein localization, yet many challenges remain to be tackled.

The work at hand describes a series of novel machine learning-based approaches that predict protein localization from amino acid sequence. Protein localization is predicted at different resolution levels: (i) a cell, (ii) a compartment and (iii) a pathogenic organism. The first approach employs machine learning (profile kernel Support Vector Machines) to predict protein sub-cellular localization. Prediction performance is made 25% better by adding homology-based inference. The improved method was made publicly available as a web server and was used to annotate over 1,000 entirely sequenced proteomes.

Predicting protein localization at a resolution of a single compartment is a harder problem due to the lack of experimental data. This work presents another method that combines homology-based inference with machine learning to predict proteins in 13 sub-nuclear localizations. In addition, a database that archives all experimentally known nuclear signals, *i.e.* “zip codes” that guide nuclear protein import and export, is described. Learned from the set of experimental signals, the database suggests a two-fold larger set of potential computationally determined signals that await their experimental verification.

Knowledge of protein localization can assist in the identification of pathogenic bacteria. The type III secretion system is a key mechanism for the transport of bacterial effector proteins directly into the cytoplasm of host cells. Similar to approaches for other localization problems, the novel method described here combines homology-based inference with machine learning to predict effector proteins. It improves up to three-fold in the prediction performance compared to the state-of-the-art. This method was also made available as a web server and was used to annotate all entirely sequenced prokaryotes.

Finally, a linked annotation resource is envisioned that could by unifying various annotations from biomedical texts complement annotations in existing biological databases.

Zusammenfassung

Zellen sind in Kompartimente unterteilt, die jeweils für eine andere zelluläre Funktion zuständig sind. Viele Proteinfunktionen können nur in bestimmten Kompartimenten ausgeübt werden. Deshalb ist die subzelluläre Lokalisierung wichtig, um die Funktion einzelner Proteine umfassend zu verstehen. Während die Anzahl bekannter Proteine stetig wächst, gestaltet sich die Proteinlokalisierung trotz des Fortschritts in experimentellen Techniken hingegen problematischer. Die Menge an Proteinen mit bekannter Lokalisierung steigt daher deutlich geringer. Viele computergestützte Methoden wurden entwickelt, um Proteinlokalisierung vorherzusagen, allerdings können die Methoden weiterhin verbessert werden.

Die vorliegende Arbeit beschreibt eine Reihe von neuen Machine Learning basierten Methoden, welche die zelluläre Lokalisierung der Proteine anhand ihrer Aminosäuresequenz vorhersagen. Dies geschieht auf den drei Ebenen: (i) Zelle, (ii) Kompartiment und (iii) pathogener Organismus. Die erste Methode verwendet Machine Learning (Profile Kernel Support Vector Maschinen), um eine Lokalisierung in verschiedenen subzellulären Kompartimenten vorherzusagen. Zusätzlich kann die Vorhersagegenauigkeit um weitere 25% verbessert werden, indem Machine Learning mit einer Homologie basierten Inferenz kombiniert wird. Die Folgemethode wurde als Web Server zur Verfügung gestellt und auf mehr als 1.000 Proteomen von derzeit sequenzierten Organismen angewendet.

Die Vorhersage von Proteinlokalisierung innerhalb eines einzigen Kompartiments stellt aufgrund des Fehlens von experimentellen Daten ein größeres Problem dar. Diese Arbeit stellt eine weitere Methode vor, die Machine Learning mit Homologie basierten Inferenz kombiniert, um Proteine in 13 sub-nukleare Kompartimente vorherzusagen. Darüber hinaus wird eine Datenbank vorgestellt, die alle experimentell bekannten Nukleuslokalisierungssignale ("Postleitzahlen") enthält. Diese steuern den Proteintransport in und aus dem Zellkern. Basierend auf experimentell bestimmten Signalen, enthält die Datenbank eine zweifach größere Anzahl von potenziellen, neu berechneten Signalen, die noch auf experimentelle Bestätigung warten.

Die Kenntnis der Proteinlokalisierung kann auch bei der Identifizierung von pathogenen Bakterien helfen. Das "Typ III-Sekretionssystem" ist ein essenzielles System für die Sekretion von bakteriellen Effektorproteinen direkt in das Zytoplasma der Wirtszellen. Ähnlich zu den Methoden für andere Lokalisierungsprobleme, kombiniert die hier beschriebene neue Methode Machine Learning mit Homologie basierten Inferenz, um Effektorproteine vorherzusagen. Die Methode verbessert die Vorhersageleistung von Effektorproteinen bis zu einem Dreifachen im Vergleich zu state-of-the-art Methoden. Die neue Methode wurde auch als Web Server zur Verfügung gestellt und auf Proteomen aller vollständig sequenzierten Prokaryoten angewendet.

Schließlich wird ein Konzept vorgestellt, das verschiedene Annotationen aus biomedizinischen Texten miteinander verknüpft und so Annotationen in existierenden biologischen Datenbanken ergänzt.

Acknowledgements

First and foremost, I would like to thank my supervisor Burkhard Rost. Throughout my years in his lab Burkhard gave me tremendous training opportunities. Thanks to Burkhard I became from early on a reviewer for a number of highly ranked journals in our field of computational biology. Burkhard's continuous support motivated me to present my work at international meetings and conferences, thus expanding my scientific network and the visibility of my research significantly. Burkhard is a visionary leader and from our many enjoyable discussions I learned how to tackle things from different perspectives, which is very valuable for me, not only with respect to research.

I am also deeply thankful to Prof. Yana Bromberg, whose lab I visited in the United States. I felt very welcomed in her group and it was a pleasure being part of it. I thank Yana for sharing with me her knowledge and her many innovative ideas. Yana has always been extremely supportive in my any endeavor. On top, she has a very pleasant personality and an excellent sense of humor.

In this context, I would like to thank the members of both labs for the working environment that I greatly enjoyed. Many thanks go to Guy Yachdav for being a friend and a mentor. I learned so much from you! Thanks also to Marlena Drabik, Inga Weise and Lothar Richter for their help with the administrative matters; to Timothy Karl and Laszlo Kajan for the help with the computer cluster; to Edda Kloppmann, Andrea Schafferhans, Tobias Hamp, Maximilian Hecht, Juan Miguel Cejuela, Chengsheng Zhu, Jonas Reeb, Esmeralda Vicedo, Arthur Dong, Thomas Hopf, Yannick Mahlich, Maximilian Miller, Christian Schaefer, Marco Punta, Marc Offman, Shaila Roessle, Maina Bitar and Dedan Githae for our many insightful scientific discussions and the fun we had together. I thank Silvana Wolf, Mohamed Ahmed and Shrikant Vinchurkar whose Master, Bachelor and Interdisciplinary projects I had an honor to supervise, and whose work became part of this dissertation.

I am thankful to Prof. Daniel Cremers and Prof. Iris Antes who agreed to be on my committee on such short notice!

My research was supported by a scholarship of the Ernst Ludwig Ehrlich Scholarship (ELES) Program. I thank Prof. Martin Vingron for accepting me in this excellent program, and Eva Lezzi and Dmitrij Belkin for being my points of contact. The discussions with other ELES scholars and the variety of offered seminars were very inspiring for me.

Lastly, I want to thank my family. My lovely mother Nelli has always motivated me to follow my dreams; she made me think I can achieve any goal set for me. My three little nephews – Jan, Katerina and Aleksandr – are my endless source of joy and happiness. Their parents, my brother Valerij and his wife Margarita, as well as Margarita's mother Ludmilla are always there for me; their support is unconditional. Finally, I want to thank my significant other, Taras Serikov, for his continuous patience, love, and everything else.

List of publications

The work at hand constitutes a cumulative dissertation. The methodologies and results presented in Chapters 2, 3 and 7 have been published in the following peer-reviewed articles (the articles are included in this dissertation):

- **Tatyana Goldberg**, Tobias Hamp and Burkhard Rost (2012). *LocTree2 predicts localization for all domains of life*. Bioinformatics, 28(18):i458-i465.
- **Tatyana Goldberg**, Maximilian Hecht, Tobias Hamp, Timothy Karl, Guy Yachdav, Nadeem Ahmed, Uwe Altermann, Philipp Angerer, Sonja Ansorge, Kinga Balasz, Michael Bernhofer, Alexander Betz, Laura Cizmadija, Kieu Trinh Do, Julia Gerke, Robert Greil, Vadim Joerdens, Maximilian Hastreiter, Katharina Hembach, Max Herzog, Maria Kalemanov, Michael Kluge, Alice Meier, Hassan Nasir, Ulrich Neumaier, Verena Prade, Jonas Reeb, Aleksandr Sorokoumov, Ilira Troshani, Susann Vorberg, Sonja Waldraff, Jonas Zierer, Henrik Nielsen and Burkhard Rost (2014). *LocTree3 prediction of localization.* Nucleic Acids Research, 42:W350-5.
- **Tatyana Goldberg**, Shrikant Vinchurkar, Juan Miguel Cejuela, Lars Juhl Jensen and Burkhard Rost (2015). *Linked annotations: a middle ground for manual curation of biomedical databases and text corpora.* BMC Proceedings, 9(Suppl 5): A4.

During the duration of the work described here I have also co-authored the following articles, which are cited in this dissertation:

- Jordan A. Ramilowski, **Tatyana Goldberg**, Jayson Harshbarger, Edda Kloppman, Marina Lizio, Venkata P. Satagopam, Masayoshi Itoh, Hideya Kawaji, Piero Carninci, Burkhard Rost and Alistair R.R. Forrest (2015) *A draft network of ligand-receptor-mediated multicellular signalling in human*. Nature Communications, 22;6:7866.
- Esmeralda Vicedo, Zofia Gasik, Yu-An Dong, **Tatyana Goldberg** and Burkhard Rost (2015). *Protein disorder reduced in Saccharomyces cerevisiae to survive heat shock.* F1000Research, 4:1222.
- Anke Graessel, Stefanie M. Hauck, Christine von Toerne, Edda Kloppmann, **Tatyana Goldberg**, Herwig Koppensteiner, Michael Schindler, Bettina Knapp, Linda Krause, Katharina Dietz, Carsten B. Schmidt-Weber and Kathrin Suttner (2015)

Combined Omics Approach to Generate the Surface Atlas of Human Naive CD4+ T Cells during Early T-Cell Receptor Activation. Molecular & Cellular Proteomics, 14(8):2085-102.

- Javad Zahiri, Mortez Mohammad-Noori, Reza Ebrahimpour, Samaneh Saadat, Joseph H. Bozorgmehr, **Tatyana Goldberg** and Ali Masoudi-Nejad (2014). *LocFuse: human protein-protein interaction prediction via classifier fusion using protein localization information*. Genomics, 104(6 Pt B):496-503.
- Guy Yachdav, Edda Kloppmann, Laszlo Kajan, Maximilian Hecht, **Tatyana Goldberg**, Tobias Hamp, Peter Hönigschmid, Andrea Schafferhans, Manfred Roos, Michael Bernhofer, Lothar Richter, Haim Ashkenazy, Marco Punta, Avner Schlessinger, Yana Bromberg, Reinhard Schneider, Gerrit Vriend, Chris Sander, Nir Ben-Tal N and Burkhard Rost B (2014). *PredictProtein--an open resource for online prediction of protein structural and functional features*. Nucleic Acids Research, 42:W337-43.
- Manuel Corpas, Rafael Jimenez, Seth J Carbon, Alex García, Leyla Garcia, **Tatyana Goldberg**, John Gomez, Alexis Kalderimis, Suzanna E Lewis, Ian Mulvany, Aleksandra Pawlik, Francis Rowland, Gustavo Salazar, Fabian Schreiber, Ian Sillitoe, William H Spooner, Anil S. Thanki, José Villaveces, Guy Yachdav and Henning Hermjakob (2014). *BioJS: an open source standard for biological visualisation - its status in 2014*. F1000Research, 13;3:55.
- Tobias Hamp, **Tatyana Goldberg** and Burkhard Rost (2013). *Accelerating the Original Profile Kernel*. PLoS One, 8(6):e68459.

List of Figures and Tables

Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 6:

Chapter 7:

Tables

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 6:

Chapter 7:

1 Introduction

1.1 Sub-cellular localization is an aspect of protein function

Compartmentalization of the cell

Prokaryotic cells are generally surrounded by a single plasma membrane (gram-negative bacteria have an additional outer membrane) that controls the flow of various substances in and out of the cell. Eukaryotic cells, in contrast, are typically much larger than that of prokaryotes and are divided into several intracellular membrane-bound compartments, called organelles, each responsible for a different cellular function. For example, the nucleus hosts the genetic material resembling a library, and the mitochondria provide the energy resembling a power plant. Therefore, proteins residing in the same sub-cellular compartment often contribute to the same cellular function.

A recent study by Bell and colleagues [1] presents potential evidence for the first forms of life to have evolved as early as 4.1 billion years ago, during the period following Earth's formation. For the next two billion years, unicellular prokaryotes (*i.e.* archaea and bacteria) presented the only form of life until 2.1 billion years ago, unicellular prokaryotes aggregated to become multicellular eukaryotes. Remarkably, the event of multicellularity occurred dozens of times independently [2] and opened completely new ways of life to become available, *i.e.* as fungi, animals and plants.

It is widely accepted that one of the critical steps in the formation of eukaryotes was the event of *endosymbiosis* – invasion of a host prokaryotic cell by a smaller prokaryotic cell [3]. For example, both mitochondria and chloroplasts originated in this way. These organelles are similar to bacteria not only in size, but also in the reproduction by diving in two. Most importantly, both organelles contain their own DNA, which replicates independently of the host cell's cycle [4]. Other eukaryotic organelles, such as the Endoplasmic reticulum (ER), Golgi apparatus, endosomes and lysosomes are believed to have evolved from the pitching off of special patches of the plasma membrane [5]. Finally, the origin of the nucleus remains unclear - whether it formed by an endosymbiont that corresponds to the nuclear compartment or by the internalization of the plasma membrane that became organized around the chromatin [6-8].

Günter Blobel wins 1999 the Nobel Prize for protein targeting system

A true revolution in the modern cell biology traces back to 1945 when Keith R. Porter, Albert Claude and Ernest F. Fullam from the Rockefeller Institute for Medical Research published the first image of a eukaryotic cell as seen with an electron microscope [9]. While earlier light microscopes allowed seeing the shape of the cell and its major compartment, the nucleus, the high resolution electron microscope allowed for the first time to see clear structures of other organelles within a cell (Figure 1). The techniques of electron microscopy were steadily improved in the next years, which in 1955 led to the identification of ribosomes (first named “Palade granules”) [10], the molecular machines responsible for the synthesis of novel proteins. Moreover, it led to the realization that different sub-cellular structures carry out different cellular functions and in order for a protein to be secreted out of the cell, it must enter a so-called secretory pathway for its transport from the cytoplasm, where it is synthesized, to the cell’s exterior traversing the plasma membrane [11-18].

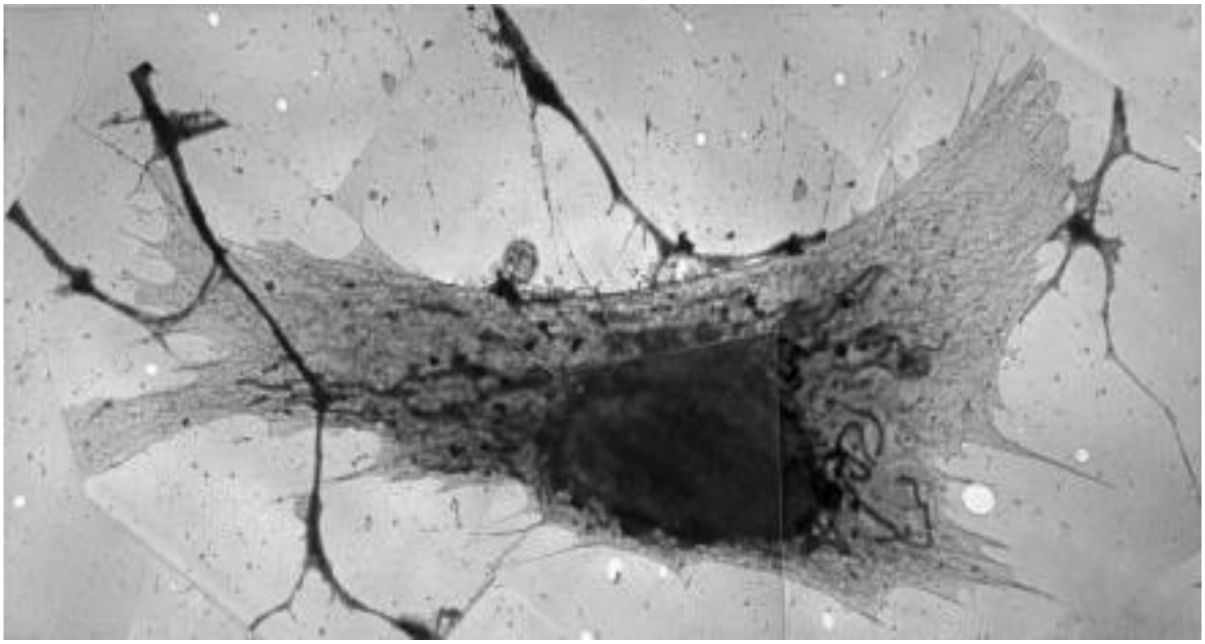


Figure 1: First published high resolution image of a eukaryotic cell. The figure shows the first electron microscope image of an intact eukaryotic cell published in 1945 by Keith R. Porter, Albert Claude, and Ernest F. Fullam [9]. The cell is a cultured fibroblast, originated from a chick embryo. Magnified 1600 times, this image reveals cell’s major sub-cellular compartments, including the nucleus, mitochondria, cytoplasm, Golgi apparatus, the extra-cellular space and a "lace-like reticulum", which Porter later named the "Endoplasmic Reticulum" [19]. Other major compartments of a eukaryotic cell, not shown here, are the chloroplasts, plastids (both in plants), lysosomes, peroxisomes and the vacuole. The image montage was taken from [20].

Günther Blobel was the first scientist who described in 1975 the mechanism of how proteins traverse cellular membranes, including those of organelles, a scientific breakthrough that was awarded with a Nobel Prize in 1999. More specifically, the Prize was awarded for the discovery that "*proteins have intrinsic signals that govern their transport and localization in the cell*" [21]. In his work, Blobel introduced a zip code-like structure of the cell, where each protein possesses an organelle-specific "address tag" or a "zip code" in its amino acid sequence that is recognized by receptors in the membrane of the targeted organelle. Upon recognition, the protein is translocated to the organelles across a channel in their membrane where it can then perform its cellular function (Figure 2). Blobel called the zip codes *signal sequences* [22] and the theory of protein transfer to the membranes of organelles, *the signal hypothesis* [23, 24]. It turned out that the protein targeting mechanism based on signal sequences, proposed by Blobel, is strongly conserved and is operating similarly across all three domains of life (*i.e.* in Archaea, Bacteria and Eukaryota) [25-29].

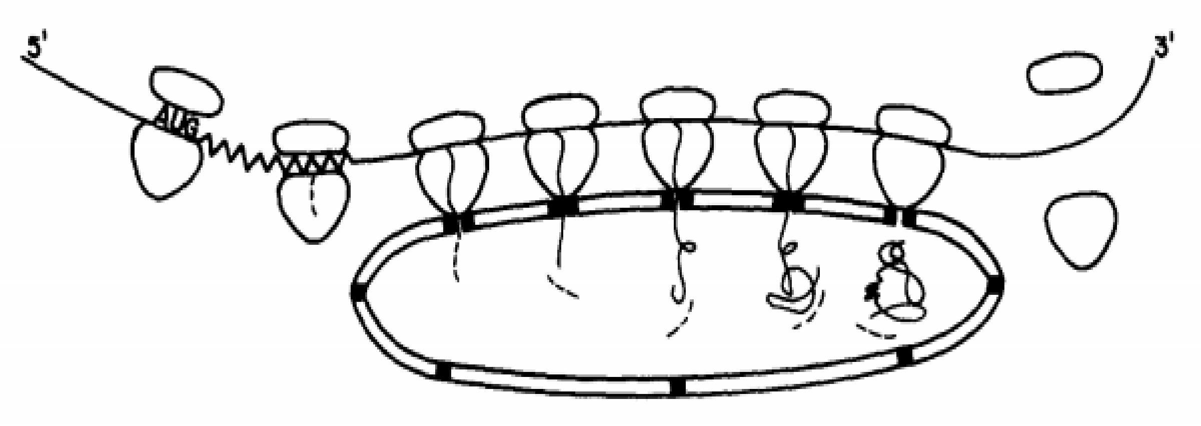


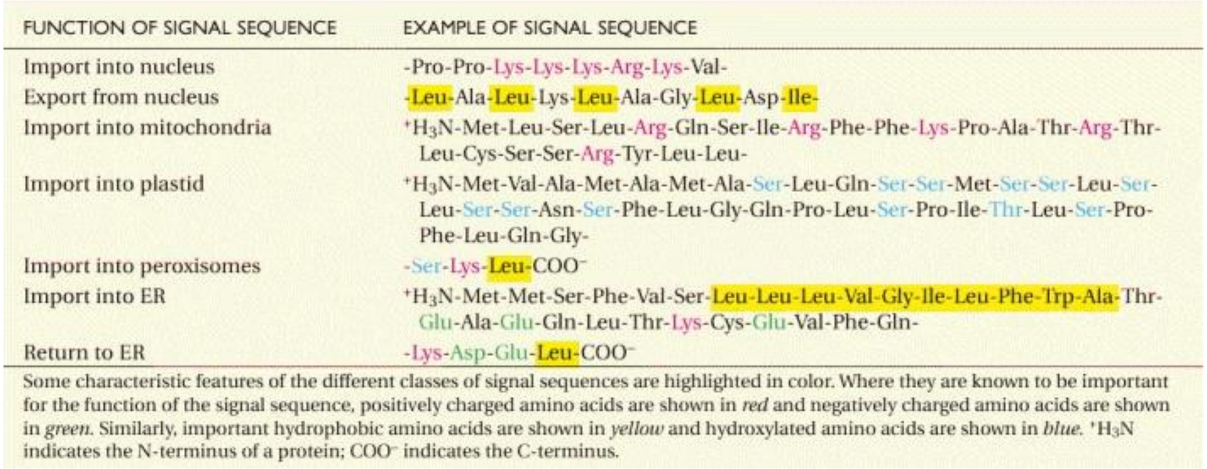
Figure 2: Blobel's signal hypothesis for the transfer of proteins across membranes. The figure illustrates the signal hypothesis introduced by Günter Blobel in 1975 [23]. A protein destined for the secretion from the cell (the mRNA encoding the protein is indicated by a long black line) is synthesized by ribosomes (white structures surrounding the mRNA) that associate with the ER. The codons in the region after the initiation AUG codon are signal codons (indicated by a zig-zag line) whose translation results in a signal sequence (indicated by a dashed line) on the N-terminus of the nascent protein. Emergence of this signal sequence triggers the attachment of the ribosome to a channel in the ER membrane, where the ever growing protein can pass through until the signal sequence is cleaved and the protein is released into the lumen of the ER. Subsequently, the protein can be transported out of the cell. The figure was taken from [23].

The vast majority of sorting signals remain unknown

On the basis of previous results, Blobel established in 1980 general principles of the cellular "protein targeting" machinery [30]. Blobel stated that amino acid sequences of the transported proteins contain topogenic organelle-specific targeting signals (or signal peptides), which are recognized by selective signal receptors that physically bind to them. The interaction between a signal and a receptor then initiates protein transport to cellular membranes and allows a protein to pass these membranes in its unfolded state. This protein translocation mechanism was shown to occur at the prokaryotic plasma membrane, and eukaryotic mitochondrial, chloroplast, thylakoid, ER and peroxisome membranes [30]. Later, Blobel also described the protein nuclear transport through nuclear pore complexes (NPCs; further described in Chapter 6), which allows proteins to pass the nuclear membrane in their folded state. This transport mechanism also requires the presence of specific targeting signals in the amino acid sequences of transported proteins [31, 32]. The importance of cellular targeting signals was shown by (i) removing them from the sequences of transported proteins, thus inhibiting their cellular transport and (ii) appending signal sequences to cytoplasmic proteins, thus mediating their transport to other sub-cellular compartments. Signal sequences are thus both necessary and sufficient for protein cellular sorting. Generally, signal sequences can be divided into two classes [5]:

- *Signal sequences*: are short stretches of consecutive residues in the amino acid sequences of transported proteins that are exposed when proteins are folded. Signal sequences usually occur at one of the ends in proteins amino acid sequences, but can also occur anywhere else in the sequences.
- *Signal patches*: are formed through amino acids that are physically separated in the sequences of transported proteins. However, once a protein folds into its three-dimensional state, the patches come together and form a signal on the surface of a folded protein.

Signal sequences can vary greatly between proteins destined for the same sub-cellular compartment (*e.g.* over 2,000 different nuclear localization signals are reported in Swiss-Prot [33]; Materials and Methods in Chapter 6). However, their physical properties, such as hydrophobicity or polarity, often seem to be more important in the signal recognition process than the exact amino acid sequence (Figure 3). Due to the lack of a consensus sequence determining a protein translocation to a certain sub-cellular compartment, it is extremely difficult to determine signal sequences experimentally and for signal patches the situation is



FUNCTION OF SIGNAL SEQUENCE	EXAMPLE OF SIGNAL SEQUENCE
Import into nucleus	-Pro-Pro-Lys-Lys-Lys-Arg-Lys-Val-
Export from nucleus	-Leu-Ala-Leu-Lys-Leu-Ala-Gly-Leu-Asp-Ile-
Import into mitochondria	^+H_3N-Met-Leu-Ser-Leu-Arg-Gln-Ser-Ile-Arg-Phe-Phe-Lys-Pro-Ala-Thr-Arg-Thr-Leu-Cys-Ser-Ser-Arg-Tyr-Leu-Leu-
Import into plastid	^+H_3N-Met-Val-Ala-Met-Ala-Met-Ala-Ser-Leu-Gln-Ser-Ser-Met-Ser-Ser-Leu-Ser-Leu-Ser-Ser-Asn-Ser-Phe-Leu-Gly-Gln-Pro-Leu-Ser-Pro-Ile-Thr-Leu-Ser-Pro-Phe-Leu-Gln-Gly-
Import into peroxisomes	-Ser-Lys-Leu-COO^-
Import into ER	^+H_3N-Met-Met-Ser-Phe-Val-Ser-Leu-Leu-Leu-Val-Gly-Ile-Leu-Phe-Trp-Ala-Thr-Glu-Ala-Glu-Gln-Leu-Thr-Lys-Cys-Glu-Val-Phe-Gln-
Return to ER	-Lys-Asp-Glu-Leu-COO^-

Some characteristic features of the different classes of signal sequences are highlighted in color. Where they are known to be important for the function of the signal sequence, positively charged amino acids are shown in *red* and negatively charged amino acids are shown in *green*. Similarly, important hydrophobic amino acids are shown in *yellow* and hydroxylated amino acids are shown in *blue*. ^+H_3N indicates the N-terminus of a protein; COO^- indicates the C-terminus.

Figure 3: Typical signal sequences involved in protein cellular sorting. The figure shows examples of signal sequences that target protein transport to different sub-cellular compartments. Typical physical properties of these sequences, which appear to be more important for protein sorting than the sequences themselves, are highlighted in color. Figure was taken from [5].

even worse, as they require knowledge about the three-dimensional structure of transported proteins. Thus, a vast majority of sorting signals remains unknown (*e.g.* Swiss-Prot nuclear localization signals can be mapped in sequences of less than 10% of all known nuclear proteins; Results and Discussion in Chapter 6). To remedy this situation, a number of computational methods have been developed that predict protein sub-cellular localization through a number of conceptually different approaches. These approaches require information other than the presence of signal sequences only.

1.2 Applications of protein sub-cellular localization data

The identification of signal peptides allowed a cell biologist for the first time to reconstruct protein transport *in vitro* and to analyze cellular functions outside of a living cell, which was nearly impossible before. This opened new possibilities of a significant impact also in clinical research, as the sub-cellular localization is essential for the protein’s functional role in a cell.

Knowledge of protein sub-cellular localization can, for example, be used in the identification of novel drug targets. Over two thirds of known drugs target proteins that are localized in the extra-cellular space and the plasma membrane [34, 35]. Proteins localized in these compartments are relatively easy to access, so that drugs targeting them do not require substantial modification. This is different for drugs that target proteins in intracellular compartments. During development, these drugs are designed in a way that they exhibit cellular sorting signals (*i.e.* zip codes) that allow them to pass through the plasma membrane and to reach their appropriate sub-cellular location [36]. For example, anticancer drugs target various nuclear proteins that are involved in *e.g.* DNA replication. To reach these proteins, drugs are often attached to specific viral machineries that lack pathogenic components but allow drug delivery into the cell nucleus [36]. Mitochondria are also often targeted in the drug therapy both in host cells and in parasites. In host cells, mitochondrial proteins can serve as anticancer targets, while in parasitic cells for example inhibiting the electron transfer chain is a successful antimicrobial intervention [36]. To reach mitochondria, a drug must contain a mitochondrial targeting signal.

Knowledge of protein sub-cellular localization can further help in understanding the molecular mechanisms of several human genetic diseases. If a sorting signal gets modified or disrupted, the protein carrying this signal can no longer reach its correct sub-cellular destination and becomes mis-localized. Aberrant protein localizations have been observed in the pathogenesis of human diseases as diverse as metabolic, cardiovascular and neurodegenerative diseases, as well as cancer [37]. One example are mutations within the nuclear localization signal of the sex-determining region Y protein (SRY), which prevent the protein from entering the nucleus and promote its mis-localization in the cell cytoplasm. The loss of nuclear function of SRY has been linked to a disease where developmental defects include male-to-female sex reversal, also known as Swyer syndrome [38]. Proteins can also mis-localize due to alterations in the elements of the protein sorting machinery. For example, dysregulations of nuclear pore complexes have been linked to the development of cardiovascular and neurodegenerative diseases [37]. Thus, the identification of disease-related protein mis-localizations offers an opportunity to normalize or interfere with the aberrant localization using therapeutic agents.

Finally, because sub-cellular localization limits interacting partners to those proteins that reside in spatially proximal or equal sub-cellular compartments, knowledge of protein localization can also be used in assessing protein-protein interaction data, such as those coming from noisy high-throughput experiments [39]. Interacting proteins are confined to particular biological processes and are likely to have similar functional annotations. Therefore, knowledge of the sub-cellular localization of a protein is also important in assigning function to its interacting partners that are yet un-annotated [40-44].

1.3 Experimental characterization of protein localization

One of the most prominent experimental techniques to determine *in vivo* steady-state sub-cellular localization of proteins is based on green fluorescent protein (GFP) tagging. GFP was originally isolated from the jellyfish *Aequorea victoria.* The protein is composed of 238 amino acid residues (27kDa) and exhibits green fluorescence when excited with blue light, without the need for any co-factors [45-47]. Therefore, any cDNA can be fused with the GFP coding sequence and the localization of the expressed GFP can be monitored using a microscope in living cells (Figure 4). Subsequently, the respective cDNAs can be extracted from cells, cloned and sequenced. This strategy of using GFP has led to a number of sub-cellular localization screening assays [48-52]. One example is the high-throughput study of the yeast proteome by Huh *et al.* [53], where over 4,000 *S. cerevisiae* proteins (representing about 60% of the whole proteome) were GFP tagged and analyzed. While the GFP-tagging method is undoubtedly powerful, it has also limitations. The GFP tag may interfere with the correct protein localization. While this interference may not apply to each and every protein, the visualization of each tagged protein is clearly a limiting factor.

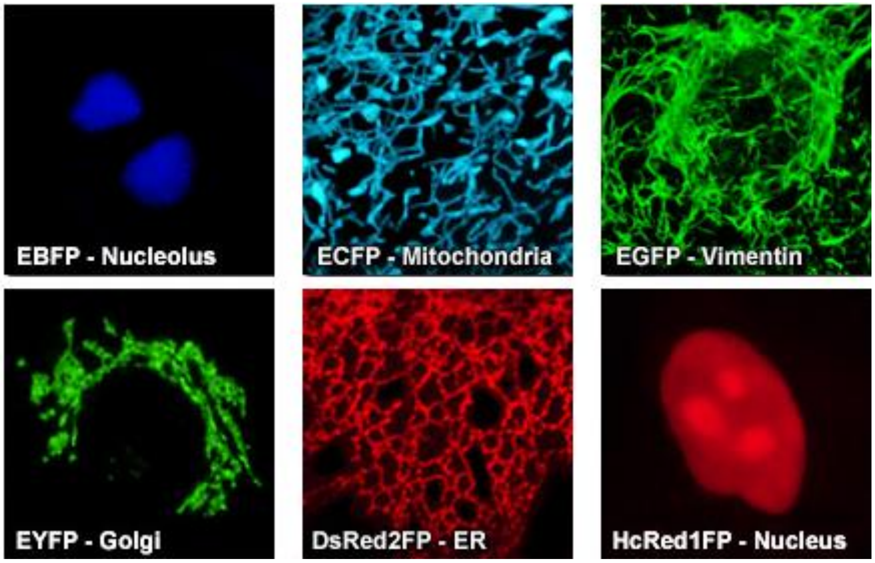


Figure 4: Fluorescent protein labelling in living cells. The figure shows fluorescence microscope images of protein markers exclusively localized to five different sub-cellular localizations (nucleolus, mitochondria, the Golgi apparatus, Endoplasmic Reticulum and nucleus) and of a protein vimentin that is known to be attached to the nucleus, Endoplasmic Reticulum and mitochondria. The fluorescent tagging was done using enhanced green fluorescent protein (EGFP) and its derivatives: blue fluorescent protein (EBFP), cyan fluorescent protein (ECFP), yellow fluorescent protein (YFP) and red fluorescent proteins (DsRed2FP and HcRed1FP). Figure was taken from [54].

In the post-genomic era, cheaper and faster solutions are needed to systematically analyze the localization of proteins in larger proteomes. The pioneering work towards the analysis of protein localization in human was done by Matthias Uhlen and colleagues [55, 56], who have generated and tested antibodies directed at 700 proteins, representing all major protein families (*e.g.* kinases, protein receptors, transcription factors and nuclear receptors). The localization of antibody-protein interactions was analyzed in nuclear, cytoplasmic and plasma membrane compartments, which was detected using GFB-based immunofluorescence. An important bottleneck for this approach, however, is the specificity and selectivity of antibodies, which need to be rigorously evaluated. Uhlen and colleagues suggest [55] to use two antibodies, best generated in different laboratories, for targeting the same gene product.

Significant advances in organelle proteomics allowed extracting entire organelles (*e.g.* the Golgi apparatus, mitochondria, lysosomes, peroxisomes, nucleus and the ER) and analyzing their proteomes [57]. Organelles purification is done through homogenization of cells and fractionation of its components (*i.e.* organelles) using a number of centrifugation techniques. Centrifugation separates the components from each other based on their size and density. Another fractionation technique that can be used with centrifugation is the isolation of cellular components using antibodies targeted at the cytoplasmic domain of an organelle transmembrane protein or a molecular tag. The proteins residing in isolated organelles can then be identified using Mass Spectrometry (MS) techniques. Though the MS-based proteomics has provided impressive results, enriching databases with proteins from various sub-cellular localizations, they have also limitations. Most importantly, they provide only a snapshot of proteins residing in an organelle at a particular time point. Also, proteins only transiently associated with an organelle are likely to be missed.

To overcome the limitations of the organellar fractionation techniques listed above, Matthias Mann and colleagues applied the approach of protein correlation profiling to map 1,404 mouse liver proteins to 10 sub-cellular localizations [58]. This approach was described in their earlier work that identified over 20 centrosomal proteins that were previously not known to be localized there [59]. First, cells were disrupted and the centrosomes were purified by centrifugation. The resulted fractions were digested with proteases and the peptides analyzed by MS. The abundance of each peptide in each fraction was determined and the abundancies were compared to the abundance of peptides from known centrosomal proteins (marker proteins). The correlation between the profiles indicated the likelihood of a protein being centrosomal (Figure 5). Thus, the major advantages of this technique are in the possibility of studying proteins localized to organelles that are difficult to purify and in not requiring antibodies or other protein tagging. All fractionation methods, however, rely on the presence of proteins within an organelle during purification. Proteins transiently attached to membranes or peripheral membrane proteins are difficult to study with these methods.

Despite the huge continuous effort in improving experimental identification techniques for protein localization, the proteomes of completely sequenced organisms remain largely un-annotated. For instance, the best studied organism yeast has less than 2/3 of its proteins annotated; for other organisms including human this number is significantly lower (discussed in Chapters 2 and 3). Therefore, bioinformatics approaches are sought to extend protein localization maps and support experimental datasets.

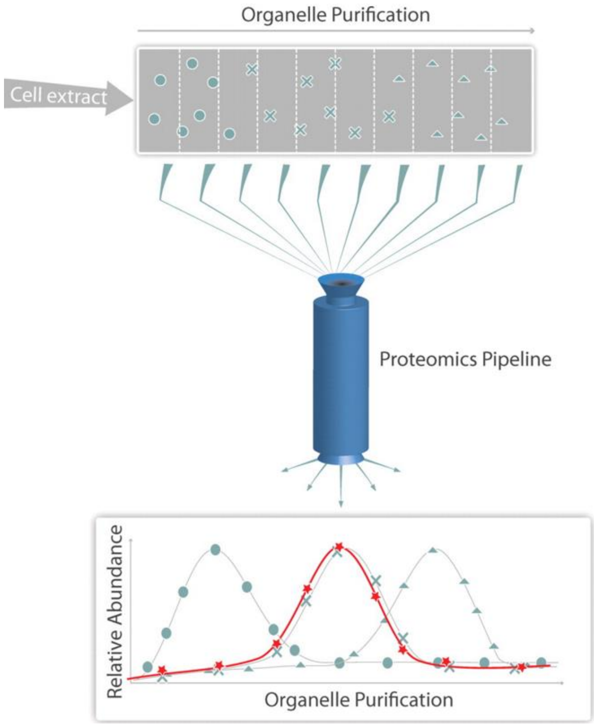


Figure 5: Workflow of the protein correlation profiling analysis. Organelles are purified from cells and divided into fractions using *e.g.* centrifugation techniques (top gray box; three types of organelles are indicated by circles, crosses and triangles). These are then subjected to proteases that break down in the organelles contained proteins into peptides, which are subsequently analyzed by a proteomics pipeline (*e.g.* a mass spectrometer). The abundance profiles of peptides across all fractions (bottom box; profiles of proteins from three organelles are given by three blue lines) are compared to the abundance profile of known marker proteins of an organelle of interest (red line). Proteins whose profiles correlate with those of marker proteins (red line - marker; line with crosses - candidate) are identified as candidates localized to an organelle, while other proteins are identified as contaminants. Figure was taken from [60].

1.4 *In silico* methods predicting protein sorting

Localization predictions are a common playground for function prediction methods

Predicting the sub-cellular localization of proteins computationally is one of the central challenges in bioinformatics. Protein localization is one aspect of protein function and in comparison to other protein functional features much more easily identifiable. Experimental studies have shown that proteins may travel between different sub-cellular compartments, yet most of them are functional within a single compartment for the largest part of their lifetime [53, 58, 61]. Furthermore, the cellular sorting mechanism is relatively well understood and experimental localization data is available in public databases for a large number of proteins. For instance, the manually annotated database Swiss-Prot [33] contains experimental localization information for more than 24,500 proteins (release 2015_12). These however constitute less than 0.05% of all known proteins (percentage is based on the UniProt [62] release 2015_12). Best computational methods have already achieved impressive levels of prediction performance [63, 64] and have been incorporated in proteome annotation pipelines to complement experiments [44, 65]. However, most of these methods were developed with the aim of predicting localization either at a specific localization site or in specific organisms.

The first published computational method that predicted protein localization from the protein amino acid sequence was PSORT, developed by Nakai and Kanehisa in 1991 [66]. Most reliable annotations however remain those that are derived from sequence homology, *i.e.* localization information is transferred from experimentally annotated protein to its un-annotated sequence homolog. For proteins with no detectable sequence homology to annotated proteins, *de novo* machine learning methods have proven to provide reliable results. Other automatic methods annotate proteins by mining biological literature and molecular biology databases. These methods however are limited to those proteins whose annotation has already been experimentally verified and published. Methods aiming at identifying features of sorting signals and using them for localization prediction have also reached remarkable levels of performance. Hybrid approaches are those methods that combine different sources of information (*e.g. de novo* predictions and sorting signal information). Finally, meta-predictors integrate various prediction methods into one; the method with the most accurate prediction is then used for the final annotation transfer. An overview of currently widely used prediction methods is provided in Table 1.

Name	Prediction feature; prediction method	Localization sites
Sequence homology–based method		
LOChom [67]	Annotated sequence homologs; PSI-BLAST [68].	10 eukaryotic and 3 bacterial sites
N-terminal sequence–based methods		
TargetP 1.1 [69]	Amino acids composition (AAC) from 100 N-terminal residues for signal peptide prediction; Neural Network (NN). Cleavage site discovered using MEME [70].	chloroplast, mitochondria (both eukaryotic), extra-cellular space (eukaryotic and bacterial)
SignalP 4.1 [71]	AAC from 70 N-terminal residues for signal peptide and cleavage site predictions.	Extra-cellular space (eukaryotic and bacterial)
EffectiveT3 [72]	Frequencies of amino acids, short peptides, and residues with certain physico-chemical properties from 25 N-terminal residues; Naïve Bayes.	Extra-cellular space (Gram-negative bacterial)
BPBAac [73]	AAC from 100 N-terminal residues; Support Vector Machine (SVM).	Extra-cellular space (Gram-negative bacterial)
Nuclear localization signals (NLS) and nuclear export signals (NES)–based methods		
PredictNLS [74]	"*In silico* mutagenesis" of known NLS.	Nuclear import (eukaryotic)
NLSstradamus [75]	AAC within NLS; Hidden Markov Models (HMMs).	Nuclear import (eukaryotic)
NESMapper [76]	AAC within NES and in 25 N-terminal and 25 C-terminal flanking residues; activity-based profile.	Nuclear export (eukaryotic)
Text mining–based methods		
LocKey [77]	"Rule library" based on Swiss-Prot keywords; M-ary classifiers.	10 eukaryotic sites
Hybrid approaches, including de novo–based methods		
LocTree [78]	Evolutionary profile-based AAC in the entire sequence, 50 N-terminal residues and three secondary structure states, as well as output of SignalP (for eukaryotes); SVMs.	5 animal, 6 plant and 3 prokaryotic sites
LocTree2 [79]	Evolutionary profile-based conservation of *k*-mers; SVMs.	18 eukaryotic, 6 bacterial and 3 archaeal sites
LocTree3 [80]	Uses PSI-BLAST homologs if available and LocTree2 otherwise.	18 eukaryotic, 6 bacterial and 3 archaeal sites
CELLO v.2.5 [81, 82]	Whole sequence-based frequencies of amino acids, di-peptides, partitioned amino acids and physico-chemical properties of amino acids; SVMs.	12 eukaryotic and 5 bacterial sites
MultiLoc2 [83]	AAC in entire sequence and N-terminal region, presence of sorting signals, phylogenetic profiles and Gene Ontology terms; SVMs.	9 animal/fungal sites and 10 plant sites
PSORTb 3.0 [84]	Sequence homologs; BLAST-P, frequent site-specific sub-sequences; SVMs, motifs and profiles derived from PROSITE [85], outer membrane motifs and transmembrane helices; HMM, signal peptides and their cleavage sites; HMM. All predictions are combined in a Bayesian network.	4 archaea/Gram-positive bacterial sites and 5 Gram-negative bacterial sites
WolFP SORT [86]	Sequence length, whole sequence-based AAC, presence of sorting signals and functional motifs, physico-chemical properties of amino acids; *k*-nearest neighbor.	12 eukaryotic sites

Table 1: Selected methods for sub-cellular localization prediction. For each method the table lists: (i) its name, (ii) features used for the prediction and the algorithm for classification (iii) predicted sub-cellular localization sites or their number and the source organism for input sequences.

Sequence homology-based methods

Homology-based inference for a protein of unknown localization U implies finding a protein with experimental localization annotation K that is sequence similar to U. This approach works, because similar sequences have similar function [87-92] and are native to the same sub-cellular localization [67]. Often, the reason for the connection between sequence similarity and the same localization is related to the same evolutionary constraints [93]. Several studies have observed a sharp conservation threshold of 50-60% sequence identity, above which pairs of proteins tend to have the same function and below which the function is different [94-96]. Other studies however indicate that these levels of sequence similarity might not be sufficient for accurate transfer of functional annotation [67, 97]. Therefore, common mistakes when searching databases for sequence homologs include: (i) using the best database hit omitting the knowledge about adequate conservation threshold for sequence similarity and (ii) ignoring the domain organization of proteins. Homology-based inferences are often used in combination with other prediction approaches [80, 98, 99]. Despite being most accurate for annotating protein sub-cellular localization, homology-based methods cannot annotate entire proteomes, as they are only applicable to proteins for which annotated homologs are available. For human they annotate 77% of the proteome, for yeast 66% and for some prokaryotes this number is lower than 1% (Chapter 3).

Sorting signal-based methods

Many methods have been developed to predict protein localization based on the identification of local sequence motifs, such as, nuclear localization and export signals for protein localization in the nucleus [100, 101] and its subsequent export [76, 102], N-terminal signal peptides for protein secretion [103, 104], or targeting peptides for localization in mitochondria and chloroplasts [105, 106] .

The first widely used method for the prediction of N-terminal sorting signals originates from the early work on secretory signal peptides of von Hejne [74, 107-110] and dates back to 1986 [111]. This method uses weight matrices, calculated from the counts of amino acids in observed signal peptides, as a linear discriminant function for the prediction of secretory proteins. The prediction accuracy for this method was reported to be 75-80%. Modern prediction methods employ machine learning algorithms, such as Neural Networks and Hidden Markov Models that learn to automatically extract correlations from the sequence data, using a set of experimentally annotated proteins as input [106, 112]. These methods boosted the prediction performance of secreted proteins to 90% accuracy. Now, it is possible to accurately predict N-terminal signals, such as secretory signals peptides [69, 71, 113, 114], mitochondrial targeting peptides [69, 114-118] and chloroplast targeting peptides [69, 114, 115, 119, 120] using machine learning-based techniques.

In contrast to N-terminal signal peptides, nuclear localization signals (NLSs) and nuclear export signals (NES) can occur anywhere in the amino acid sequence [5, 121]. Nuclear signals can be very diverse in the amino acids composition, but in general, NLS have an abundance of positively charged residues [21] and NES of hydrophobic residues [76]. One of the first attempts to predict NLS was done by Cokol and Rost [100], who successfully applied "*in silico* mutagenesis" approach to predict over 200 novel NLS. Later, several methods have been developed to predict NLS and NES using machine learning approaches that use information extracted from the signal sequence [75, 76, 122, 123].

All sorting signals-based methods are limited to those signal sequences that have already been experimentally verified. The majority of sorting signals however remains yet unknown and for signal patches the situation is even worse. Moreover, the presence of secretory peptides does not always guarantee protein secretion, as many proteins with a signal peptide are retained in the Golgi apparatus, the ER or in vesicles [5]. Alternatively, many secreted proteins use alternative pathways to cross and exit the cell [124-126].

Text mining-based methods

Before functional annotation can make an entry in a biological database, it needs to be manually extracted by an expert from the corresponding publication. These publications are stored in a public knowledgebase, such as PubMed [127]. Currently, PubMed stores over 25 million entries of biomedical literature and 500.000 new entries are added to the database each year [128]. This enormous source of knowledge is used by automatic text mining methods that extract protein localization information from the abstracts and full texts of published articles. All identified gene/protein and organism names, as well as localization occurrences need to be mapped to a controlled vocabulary or ontology, such as for example Gene Ontology [129] terms for localizations and UniProt identifiers for proteins. Such mapping presents one of the largest bottlenecks hampering the prediction performance of text mining-based methods. The evaluation of text mining methods is done on manually annotated corpora, such as GENIA for protein names and localization terms [130]. Promising results have been obtained by methods that analyze the impact of GO term comentions in texts with a Support Vector Machine (SVM) [131] classifier [132-135]. Other methods combine information from both biomedical texts and protein sequences [130, 133, 136, 137].

Other text mining-based approaches explore the functional annotation provided in UniProt [62], especially the keyword annotation. Currently, UniProt lists over 1,800 different keywords, which are organized in a controlled vocabulary of a hierarchical structure (UniProt release 2015_12). Several methods have been developed that extract rules from keywords by using machine learning methods like probabilistic Bayesian models [138], C4.5 decision trees [139, 140] and M-ary classifiers (*e.g. k*-nearest neighbor [141] and linear least-square fit [142]) [77].

De novo prediction methods

The most universal prediction methods are *de novo* methods, as they use no other information than that encoded in the protein amino acid sequence for their prediction. *De novo* methods can be applied to virtually any existing protein sequence. The fist *de novo* prediction method was developed by Nishikawa and Ooi who classified intra- and extracellular proteins based on the composition of their amino acids [143]. The success of this method is intuitively obvious – each sub-cellular compartment is characterized by its specific physico-chemical properties, so proteins localized to this compartment must evolve a different surface in order to adapt to this environment. Indeed, a correlation between protein amino acid composition and its localization has been shown by Andrade and colleagues [144]. This finding let to the development of a battery of prediction methods that exploit protein surface composition in combination with standard statistical methods [66] and machine learning techniques such as neural networks [145]. Because biological data are often small and noisy and SVMs are good at dealing with such data [146], SVMs have been shown to outperform neural networks-based methods [147]. Later developed methods incorporated information about composition of di-peptides [148] and *n*-peptides [149]. The LocTree method, developed by Nair and Rost in 2005 [78], incorporated a number of SVMs organized in a binary decision tree that resembled cellular protein sorting. It used amino acid composition in the entire sequence, the N-terminal region and in three secondary structure states; the composition was derived from evolutionary profiles. LocTree outperformed all other methods in the prediction performance.

1.5 Overview of this work

Proteins are cellular workhorses involved in nearly all processes that make life. Living cells are divided into specific sub-cellular compartments, each responsible for a different cellular function. The identification of protein localization within a cell can help in elucidating its function, as certain functions can only be performed in certain environments. Immense resources have been spent on experimentally unraveling the sub-cellular localization of proteins. However, the localization remains experimentally uncharacterized for most proteins. This calls for *in silico* methods to fill in the gap.

In Chapter 2, I describe LocTree2, a machine learning-based method for predicting protein sub-cellular localization that uses new data and lessons learned from other predictors published over the last two decades. LocTree2 classifies proteins from all three domains of life (*i.e.* Archaea, Bacteria and Eukaryota) in the so far largest number of sub-cellular localization compartments. The method outperforms existing resources and performs well even when triggered with incomplete and erroneous data.

In Chapter 3, I present LocTree3, an improvement of LocTree2 by remarkable 25 percentage points in the prediction performance. The improvement is done through a simple trick that combines homology-based inference with machine learning. For a query protein, LocTree3 first identifies a sequence homolog in the database of experimentally annotated proteins. If a homolog is available, its annotation is transferred to the query protein. Otherwise, LocTree2 is triggered for a *de novo* prediction of sub-cellular localization.

In Chapter 4, I describe LocNuclei, a predictor for protein localization at even more detailed, higher resolution level for nuclear proteins. The nucleus is a very dynamic compartment consisting of various areas, each responsible for a different function and thus hosting a different set of proteins. Experimental sub-nuclear annotations are challenging. LocNuclei is a method that inspired by LocTree3's success combines homology-based inference with machine learning to accurately predict proteins in 13 different sub-nuclear compartments. I used LocNuclei to annotate the entire human proteome.

In Chapter 5, I aim at the discovery of the so-called nuclear localization signals (NLS) and nuclear export signals (NES) that are short stretches in the amino acid sequences of nuclear proteins. They can be imagined to be “zip code” signals that help in shuttling proteins from the cytoplasm into the nucleus (NLS) and from the nucleus back into the cytoplasm (NES). In this work, I again built upon resources and ideas from many other groups and increased the set of experimentally known signals by almost an order of magnitude by reliable potential signals that await experimental verification.

In Chapter 6, I present pEffect, a method that challenges the objective of predicting pathogenic bacteria from protein sequences. The key to the success lies here again in the combination of zip code-like signals with homology-based inference and machine learning. The so-called “type III secretion system” is a pivotal mechanism for the transport of pathogenic bacterial proteins (so-called “effectors”) into the targeted host cells. Bacteria inject their effectors into targeted cells, which during infection convert host resources to work to bacterial advantage. pEffect is a method that improves up to 3-fold over the state-of-the-art. Importantly, it also sheds new light on the mechanism of effector secretion.

In Chapter 7, I discuss a “linked annotation resource”, which is an open forum for convenient collaborations between annotators of biomedical texts. Every important scientific discovery is published. Many groups put tremendous effort in mining biomedical literature to extract structured protein/gene annotations from largely unstructured texts. However, the way of sharing valuable resources still remains at a primitive level (*e.g.* through exchange of archived files). An open forum, in contrast, allows collecting annotations of various types (*e.g.* sub-cellular localization, binding sites, and effects of amino acid substitutions), linking them and making publicly available online. On the use case of protein localization I show that linked annotations can also significantly complement biological database annotations.

Finally, I present the main findings and conclusions of this work.

1.6 References

1. Bell EA, Boehnke P, Harrison TM, Mao WL: **Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon.** *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112:**14518-14521.
2. Niklas KJ, Newman SA: **The origins of multicellular organisms.** *Evolution & development* 2013, **15:**41-52.
3. Zimorski V, Ku C, Martin WF, Gould SB: **Endosymbiotic theory for organelle origins.** *Current opinion in microbiology* 2014, **22:**38-48.
4. G.M. C: *The Cell, a molecular approach. .* Boston University: Sunderland (MA): Sinauer Associates; 2000.
5. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell.* New York: Garland Science; 2002.
6. Baum DA, Baum B: **An inside-out origin for the eukaryotic cell.** *BMC biology* 2014, **12:**76.
7. Martin W: **Archaebacteria (Archaea) and the origin of the eukaryotic nucleus.** *Current opinion in microbiology* 2005, **8:**630-637.
8. Lopez-Garcia P, Moreira D: **Open Questions on the Origin of Eukaryotes.** *Trends in ecology & evolution* 2015, **30:**697-708.
9. Porter KR, Claude A, Fullam EF: **A Study of Tissue Culture Cells by Electron Microscopy : Methods and Preliminary Observations.** *The Journal of experimental medicine* 1945, **81:**233-246.
10. Palade GE: **A small particulate component of the cytoplasm.** *The Journal of biophysical and biochemical cytology* 1955, **1:**59-68.
11. Caro LG, Palade GE: **[The role of the Golgi apparatus in the secretory process. Autoradiographic study].** *Comptes rendus des seances de la Societe de biologie et de ses filiales* 1961, **155:**1750-1762.
12. Jamieson JD, Palade GE: **Role of the Golgi complex in the intracellular transport of secretory proteins.** *Proceedings of the National Academy of Sciences of the United States of America* 1966, **55:**424-431.
13. Palade GE: **The secretory cycle of the pancreatic exocrine cell.** *Kaibogaku zasshi Journal of anatomy* 1966, **41:**337-338.
14. Jamieson JD, Palade GE: **Intracellular transport of secretory proteins in the pancreatic exocrine cell. I. Role of the peripheral elements of the Golgi complex.** *The Journal of cell biology* 1967, **34:**577-596.
15. Jamieson JD, Palade GE: **Intracellular transport of secretory proteins in the pancreatic exocrine cell. II. Transport to condensing vacuoles and zymogen granules.** *The Journal of cell biology* 1967, **34:**597-615.
16. Jamieson JD, Palade GE: **Intracellular transport of secretory proteins in the pancreatic exocrine cell. 3. Dissociation of intracellular transport from protein synthesis.** *The Journal of cell biology* 1968, **39:**580-588.
17. Jamieson JD, Palade GE: **Intracellular transport of secretory proteins in the pancreatic exocrine cell. IV. Metabolic requirements.** *The Journal of cell biology* 1968, **39:**589-603.
18. Jamieson JD, Palade GE: **Synthesis, intracellular transport, and discharge of secretory proteins in stimulated pancreatic exocrine cells.** *The Journal of cell biology* 1971, **50:**135-158.
19. Porter KR, Thompson HP: **Some Morphological Features of Cultured Rat Sarcoma Cells as Revealed by the Electron Microscope.** *Cancer Research* 1947, **7:**431–438.

20. http://www.biologicalelectronmicroscopy.com/a-brief-history-of-micorscopy.html.
21. http://www.nobelprize.org/nobel_prizes/medicine/laureates/1999/.
22. Blobel G, Sabatini DD: **Ribosome-membrane interaction in eukaryotic cells.** *Biomembranes* 1971, **2:**193-195.
23. Blobel G, Dobberstein B: **Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma.** *The Journal of cell biology* 1975, **67:**835-851.
24. Blobel G, Dobberstein B: **Transfer of proteins across membranes. II. Reconstitution of functional rough microsomes from heterologous components.** *The Journal of cell biology* 1975, **67:**852-862.
25. Ngsee JK, Hansen W, Walter P, Smith M: **Cassette mutagenic analysis of the yeast invertase signal peptide: effects on protein translocation.** *Molecular and cellular biology* 1989, **9:**3400-3410.
26. Meyer DI: **The signal hypothesis — a working model.** *Trends in Biochemical Sciences* 1982, **7:**320 - 321.
27. Emr SD, Hall MN, Silhavy TJ: **A mechanism of protein localization: the signal hypothesis and bacteria.** *The Journal of cell biology* 1980, **86:**701-711.
28. Tuteja R: **Type I signal peptidase: an overview.** *Archives of biochemistry and biophysics* 2005, **441:**107-111.
29. Blobel G, Walter P, Chang CN, Goldman BM, Erickson AH, Lingappa VR: **Translocation of proteins across membranes: the signal hypothesis and beyond.** *Symposia of the Society for Experimental Biology* 1979, **33:**9-36.
30. Blobel G: **Intracellular protein topogenesis.** *Proceedings of the National Academy of Sciences of the United States of America* 1980, **77:**1496-1500.
31. Floer M, Blobel G: **The nuclear transport factor karyopherin beta binds stoichiometrically to Ran-GTP and inhibits the Ran GTPase activating protein.** *The Journal of biological chemistry* 1996, **271:**5313-5316.
32. Chook YM, Blobel G: **Karyopherins and nuclear import.** *Current opinion in structural biology* 2001, **11:**703-715.
33. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic acids research* 2000, **28:**45-48.
34. Bakheet TM, Doig AJ: **Properties and identification of human protein drug targets.** *Bioinformatics* 2009, **25:**451-457.
35. Chacinska A, Koehler CM, Milenkovic D, Lithgow T, Pfanner N: **Importing mitochondrial proteins: machineries and mechanisms.** *Cell* 2009, **138:**628-644.
36. Rajendran L, Knolker HJ, Simons K: **Subcellular targeting strategies for drug design and delivery.** *Nature reviews Drug discovery* 2010, **9:**29-42.
37. Hung MC, Link W: **Protein localization in disease and therapy.** *Journal of cell science* 2011, **124:**3381-3392.
38. McLane LM, Corbett AH: **Nuclear localization signals and human disease.** *IUBMB life* 2009, **61:**697-706.
39. Hamp T, Rost B: **Evolutionary profiles improve protein-protein interaction prediction from sequence.** *Bioinformatics* 2015, **31:**1945-1950.
40. Ofran Y, Yachdav G, Mozes E, Soong TT, Nair R, Rost B: **Create and assess protein networks through molecular characteristics of individual proteins.** *Bioinformatics* 2006, **22:**e402-407.

41. Zahiri J, Mohammad-Noori M, Ebrahimpour R, Saadat S, Bozorgmehr JH, Goldberg T, Masoudi-Nejad A: **LocFuse: human protein-protein interaction prediction via classifier fusion using protein localization information.** *Genomics* 2014, **104:**496-503.
42. Rao VS, Srinivas K, Sujini GN, Kumar GN: **Protein-protein interaction detection: methods and analysis.** *International journal of proteomics* 2014, **2014:**147648.
43. Legrain P, Wojcik J, Gauthier JM: **Protein--protein interaction maps: a lead towards cellular functions.** *Trends in genetics : TIG* 2001, **17:**346-352.
44. Ramilowski JA, Goldberg T, Harshbarger J, Kloppman E, Lizio M, Satagopam VP, Itoh M, Kawaji H, Carninci P, Rost B, Forrest AR: **A draft network of ligand-receptor-mediated multicellular signalling in human.** *Nature communications* 2015, **6:**7866.
45. Prendergast FG, Mann KG: **Chemical and physical properties of aequorin and the green fluorescent protein isolated from Aequorea forskalea.** *Biochemistry* 1978, **17:**3448-3453.
46. Tsien RY: **The green fluorescent protein.** *Annual review of biochemistry* 1998, **67:**509-544.
47. Heim R, Cubitt AB, Tsien RY: **Improved green fluorescence.** *Nature* 1995, **373:**663-664.
48. Phillips GJ: **Green fluorescent protein--a bright idea for the study of bacterial protein localization.** *FEMS microbiology letters* 2001, **204:**9-18.
49. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC: **Green fluorescent protein as a marker for gene expression.** *Science* 1994, **263:**802-805.
50. Bialkowska A, Zhang XY, Reiser J: **Improved tagging strategy for protein identification in mammalian cells.** *BMC genomics* 2005, **6:**113.
51. Simpson JC, Pepperkok R: **Localizing the proteome.** *Genome biology* 2003, **4:**240.
52. O'Rourke NA, Meyer T, Chandy G: **Protein localization studies in the age of 'Omics'.** *Current opinion in chemical biology* 2005, **9:**82-87.
53. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.
54. http://www.microscopyu.com/articles/livecellimaging/fpimaging.html.
55. Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, Andersen E, Andersson AC, Angelidou P, Asplund A, Asplund C, et al: **A human protein atlas for normal and cancer tissues based on antibody proteomics.** *Molecular & cellular proteomics : MCP* 2005, **4:**1920-1932.
56. Berglund L, Bjorling E, Oksvold P, Fagerberg L, Asplund A, Szigyarto CA, Persson A, Ottosson J, Wernerus H, Nilsson P, et al: **A genecentric Human Protein Atlas for expression profiles based on antibodies.** *Molecular & cellular proteomics : MCP* 2008, **7:**2019-2027.
57. Yates JR, 3rd, Gilchrist A, Howell KE, Bergeron JJ: **Proteomics of organelles and large cellular structures.** *Nature reviews Molecular cell biology* 2005, **6:**702-714.
58. Foster LJ, de Hoog CL, Zhang Y, Xie X, Mootha VK, Mann M: **A mammalian organelle map by protein correlation profiling.** *Cell* 2006, **125:**187-199.
59. Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M: **Proteomic characterization of the human centrosome by protein correlation profiling.** *Nature* 2003, **426:**570-574.
60. Walther TC, Mann M: **Mass spectrometry-based proteomics in cell biology.** *The Journal of cell biology* 2010, **190:**491-500.
61. Li S, Ehrhardt DW, Rhee SY: **Systematic analysis of Arabidopsis organelles and a protein localization database for facilitating fluorescent tagging of full-length Arabidopsis proteins.** *Plant physiology* 2006, **141:**527-539.
62. **UniProt: a hub for protein information.** *Nucleic acids research* 2015, **43:**D204-212.
63. Gardy JL, Brinkman FS: **Methods for predicting bacterial protein subcellular localization.** *Nature reviews Microbiology* 2006, **4:**741-751.

64. Hu Y, Lehrach H, Janitz M: **Comparative analysis of an experimental subcellular protein localization assay and in silico prediction methods.** *Journal of molecular histology* 2009, **40:**343-352.

65. Graessel A, Hauck SM, von Toerne C, Kloppmann E, Goldberg T, Koppensteiner H, Schindler M, Knapp B, Krause L, Dietz K, et al: **A Combined Omics Approach to Generate the Surface Atlas of Human Naive CD4+ T Cells during Early T-Cell Receptor Activation.** *Molecular & cellular proteomics : MCP* 2015, **14:**2085-2102.

66. Nakai K, Kanehisa M: **Expert system for predicting protein localization sites in gram-negative bacteria.** *Proteins* 1991, **11:**95-110.

67. Nair R, Rost B: **Sequence conserved for subcellular localization.** *Protein science : a publication of the Protein Society* 2002, **11:**2836-2847.

68. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic acids research* 1997, **25:**3389-3402.

69. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools.** *Nature protocols* 2007, **2:**953-971.

70. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* 1994, **2:**28-36.

71. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions.** *Nature methods* 2011, **8:**785-786.

72. Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T: **Sequence-based prediction of type III secreted proteins.** *PLoS pathogens* 2009, **5:**e1000376.

73. Wang Y, Zhang Q, Sun MA, Guo D: **High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles.** *Bioinformatics* 2011, **27:**777-784.

74. von Heijne G: **Signal sequences are not uniformly hydrophobic.** *Journal of molecular biology* 1982, **159:**537-541.

75. Nguyen Ba AN, Pogoutse A, Provart N, Moses AM: **NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction.** *BMC bioinformatics* 2009, **10:**202.

76. Kosugi S, Yanagawa H, Terauchi R, Tabata S: **NESmapper: accurate prediction of leucine-rich nuclear export signals using activity-based profiles.** *PLoS computational biology* 2014, **10:**e1003841.

77. Nair R, Rost B: **Inferring sub-cellular localization through automated lexical analysis.** *Bioinformatics* 2002, **18 Suppl 1:**S78-86.

78. Nair R, Rost B: **Mimicking cellular sorting improves prediction of subcellular localization.** *Journal of molecular biology* 2005, **348:**85-100.

79. Goldberg T, Hamp T, Rost B: **LocTree2 predicts localization for all domains of life.** *Bioinformatics* 2012, **28:**i458-i465.

80. Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, Altermann U, Angerer P, Ansorge S, Balasz K, et al: **LocTree3 prediction of localization.** *Nucleic acids research* 2014, **42:**W350-355.

81. Yu CS, Lin CJ, Hwang JK: **Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions.** *Protein science : a publication of the Protein Society* 2004, **13:**1402-1406.

82. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization.** *Proteins* 2006, **64:**643-651.

83. Blum T, Briesemeister S, Kohlbacher O: **MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction.** *BMC bioinformatics* 2009, **10:**274.

84. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26:**1608-1615.

85. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation.** *Nucleic acids research* 2010, **38:**D161-166.

86. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor.** *Nucleic acids research* 2007, **35:**W585-587.

87. Whisstock JC, Lesk AM: **Prediction of protein function from protein sequence and structure.** *Quarterly reviews of biophysics* 2003, **36:**307-340.

88. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y: **Automatic prediction of protein function.** *Cellular and molecular life sciences : CMLS* 2003, **60:**2637-2650.

89. Tamames J, Ouzounis C, Casari G, Sander C, Valencia A: **EUCLID: automatic classification of proteins in functional classes by their database annotations.** *Bioinformatics* 1998, **14:**542-543.

90. Piovesan D, Giollo M, Leonardi E, Ferrari C, Tosatto SC: **INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity.** *Nucleic acids research* 2015, **43:**W134-140.

91. Pearson WR: **Protein Function Prediction: Problems and Pitfalls.** *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2015, **51:**4 12 11-18.

92. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, et al: **A large-scale evaluation of computational protein function prediction.** *Nature methods* 2013, **10:**221-227.

93. Thornton JM, Orengo CA, Todd AE, Pearl FM: **Protein folds, functions and evolution.** *Journal of molecular biology* 1999, **293:**333-342.

94. Orengo CA, Todd AE, Thornton JM: **From protein structure to function.** *Current opinion in structural biology* 1999, **9:**374-382.

95. Wilson CA, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.** *Journal of molecular biology* 2000, **297:**233-249.

96. Pawlowski K, Godzik A: **Surface map comparison: studying function diversity of homologous proteins.** *Journal of molecular biology* 2001, **309:**793-806.

97. Rost B: **Enzyme function less conserved than anticipated.** *Journal of molecular biology* 2002, **318:**595-608.

98. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS: **PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria.** *Nucleic acids research* 2003, **31:**3613-3617.

99. Wrzeszczynski KO, Rost B: **Annotating proteins from endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes.** *Cellular and molecular life sciences : CMLS* 2004, **61:**1341-1353.

100. Cokol M, Nair R, Rost B: **Finding nuclear localization signals.** *EMBO reports* 2000, **1:**411-415.

101. Nakai K: **Protein sorting signals and prediction of subcellular localization.** *Advances in protein chemistry* 2000, **54:**277-344.

102. la Cour T, Kiemer L, Molgaard A, Gupta R, Skriver K, Brunak S: **Analysis and prediction of leucine-rich nuclear export signals.** *Protein engineering, design & selection : PEDS* 2004, **17:**527-536.

103. von Heijne G: **Protein sorting signals: simple peptides with complex functions.** *Exs* 1995, **73:**67-76.

104. Nakai K, Horton P: **PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization.** *Trends Biochem Sci* 1999, **24:**34-36.

105. Claros MG, Brunak S, von Heijne G: **Prediction of N-terminal protein sorting signals.** *Current opinion in structural biology* 1997, **7:**394-398.

106. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *Journal of molecular biology* 2000, **300:**1005-1016.

107. von Heijne G: **On the hydrophobic nature of signal sequences.** *European journal of biochemistry / FEBS* 1981, **116:**419-422.

108. von Heijne G: **Patterns of amino acids near signal-sequence cleavage sites.** *European journal of biochemistry / FEBS* 1983, **133:**17-21.

109. von Heijne G: **Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implications for protein export in prokaryotic and eukaryotic cells.** *The EMBO journal* 1984, **3:**2315-2318.

110. von Heijne G: **Signal sequences. The limits of variation.** *Journal of molecular biology* 1985, **184:**99-105.

111. von Heijne G: **A new method for predicting signal sequence cleavage sites.** *Nucleic acids research* 1986, **14:**4683-4690.

112. Nielsen H, Brunak S, von Heijne G: **Machine learning approaches for the prediction of signal peptides and other protein sorting signals.** *Protein engineering* 1999, **12:**3-9.

113. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method.** *Journal of molecular biology* 2004, **338:**1027-1036.

114. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S: **Extensive feature detection of N-terminal protein sorting signals.** *Bioinformatics* 2002, **18:**298-305.

115. Emanuelsson O, von Heijne G: **Prediction of organellar targeting signals.** *Biochimica et biophysica acta* 2001, **1541:**114-119.

116. Fujiwara Y, Asogawa M, Nakai K: **Prediction of Mitochondrial Targeting Signals Using Hidden Markov Model.** *Genome informatics Workshop on Genome Informatics* 1997, **8:**53-60.

117. Fukasawa Y, Tsuji J, Fu SC, Tomii K, Horton P, Imai K: **MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites.** *Molecular & cellular proteomics : MCP* 2015, **14:**1113-1126.

118. Savojardo C, Martelli PL, Fariselli P, Casadio R: **TPpred2: improving the prediction of mitochondrial targeting peptide cleavage sites by exploiting sequence motifs.** *Bioinformatics* 2014, **30:**2973-2974.

119. Emanuelsson O, Nielsen H, von Heijne G: **ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites.** *Protein science : a publication of the Protein Society* 1999, **8:**978-984.

120. https://urgi.versailles.inra.fr/predotar/predotar.html.

121. LaCasse EC, Lefebvre YA: **Nuclear localization signals overlap DNA- or RNA-binding domains in nucleic acid-binding proteins.** *Nucleic acids research* 1995, **23:**1647-1656.

122. Fu SC, Imai K, Horton P: **Prediction of leucine-rich nuclear export signal containing proteins with NESsential.** *Nucleic acids research* 2011, **39:**e111.

123. Lin JR, Hu J: **SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring.** *PloS one* 2013, **8:**e76864.

124. Bendtsen JD, Kiemer L, Fausboll A, Brunak S: **Non-classical protein secretion in bacteria.** *BMC microbiology* 2005, **5:**58.

125. Nickel W: **The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes.** *European journal of biochemistry / FEBS* 2003, **270:**2109-2119.

126. Stein KR, Giardina BJ, Chiang H: **The Non-classical Pathway is the Major Pathway to Secrete Proteins in Saccharomyces cerevisiae.** *Clin Exp Pharmacol* 2014, **4:155**.

127. http://www.ncbi.nlm.nih.gov/pubmed.

128. Lu Z: **PubMed and beyond: a survey of web tools for searching biomedical literature.** *Database : the journal of biological databases and curation* 2011.

129. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25:**25-29.

130. Chun HW, Yamasaki C, Saichi N, Tanaka M, Hishiki T, Imanishi T, Gojobori T, Kim JD, Tsujii J, Takagi T: **Prediction of Protein Sub-cellular Localization using Information from Texts and Sequences.** *BioNLP 2008: Current Trends in Biomedical Natural Language Processing* 2008**:**90–91.

131. Cortes C, Vapnik V: **Support-Vector Networks.** *Machine Learning* 1995, **20:**273-297.

132. Funk CS, Kahanda I, Ben-Hur A, Verspoor KM: **Evaluating a variety of text-mined features for automatic protein function prediction with GOstruct.** *Journal of biomedical semantics* 2015, **6:**9.

133. Hoglund A, Blum T, Brady S, Donnes P, Miguel JS, Rocheford M, Kohlbacher O, Shatkay H: **Significantly improved prediction of subcellular localization by integrating text and protein sequence data.** *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2006**:**16-27.

134. Fyshe A, Liu Y, Szafron D, Greiner R, Lu P: **Improving subcellular localization prediction using text classification and the gene ontology.** *Bioinformatics* 2008, **24:**2512-2517.

135. Liu Y, Guo Z, Kondrak G: **Protein Subcellular Localization Extraction and Prediction from PubMed Abstracts.** *BioKDD* 2010.

136. Shatkay H, Hoglund A, Brady S, Blum T, Donnes P, Kohlbacher O: **SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data.** *Bioinformatics* 2007, **23:**1410-1417.

137. Stapley BJ, Kelley LA, Sternberg MJ: **Predicting the sub-cellular location of proteins from text using support vector machines.** *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2002**:**374-385.

138. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R: **Predicting subcellular localization of proteins using machine-learned classifiers.** *Bioinformatics* 2004, **20:**547-556.

139. Quinlan JR: **Programs for machine learning.** *Morgan Kaufmann, San Francisco, CA* 1993.

140. Kretschmann E, Fleischmann W, Apweiler R: **Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT.** *Bioinformatics* 2001, **17:**920-926.

141. Dasarathy BV: *Nearest neighbor (NN) norms.* IEEE Computer Society Press; 1991.

142. Miller SJ: *The Method of Least Squares.* Brown University; 2006.

143. Nishikawa K, Ooi T: **Correlation of the amino acid composition of a protein to its structural and biological characters.** *Journal of biochemistry* 1982, **91:**1821-1824.

144. Andrade MA, O'Donoghue SI, Rost B: **Adaptation of protein surfaces to subcellular location.** *Journal of molecular biology* 1998, **276:**517-525.

145. Reinhardt A, Hubbard T: **Using neural networks for prediction of the subcellular location of proteins.** *Nucleic acids research* 1998, **26:**2230-2236.

146. Vapnik V: *The nature of statistical learning theory.* Information Science and Statistics; 1995.

147. Hua S, Sun Z: **Support vector machine approach for protein subcellular localization prediction.** *Bioinformatics* 2001, **17:**721-728.

148. Park KJ, Kanehisa M: **Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs.** *Bioinformatics* 2003, **19:**1656-1663.

149. Cai YD, Liu XJ, Xu XB, Chou KC: **Support vector machines for the classification and prediction of beta-turn types.** *Journal of peptide science : an official publication of the European Peptide Society* 2002, **8:**297-301.

2 LocTree2: prediction of protein cellular sorting in all domains of life

2.1 Preface

The knowledge of protein sorting within a cell can help in understanding protein function, as certain functions can only be performed in certain environments [1-3]. Though some proteins can localize in multiple compartments, most of them are functional within a single compartment [4-6]. Due to the sub-cellular localization being an easily definable functional feature, many *in silico* methods have been developed that predict localization [7-18].

In this publication, we present a novel method LocTree2 that predicts protein localization and addresses several shortcomings of the existing approaches. Namely, the method presents a common framework for all proteins in all domains of life that requires only the amino acid sequence as input. It accurately classifies proteins in the so far largest number of cellular localization classes: 18 classes for eukaryota, 6 for bacteria and 3 for archaea. It distinguishes between integral trans-membrane and water-soluble globular proteins as good as the best expert methods developed explicitly for this task [19, 20]. Even when tested on erroneous and incomplete sequence data, the method reaches high levels of performance. Similar to LocTree [13], our method implements a decision tree of localization classes imitating the protein sorting mechanism of the cell. Different from LocTree, we make binary decisions at all levels of the tree by searching through proteins of annotated localization classes with short stretches of *k* consecutive residues, *i.e.* potential localization motifs. As a proof of principle, we investigate some of the *k*-mers, which are crucial for protein classification, to be Endoplasmic Reticulum-associated. When compared to other methods, LocTree2 shows an improved prediction performance on almost all data sets tested. As suggested by one of our anonymous reviewers, we re-trained LocTree2 on old data (from year 2005) to show the improvement of our method originating from the underlying method. Indeed, the data set had only little effect on LocTree2's performance.

The study design and methodology was conceived by me and Burkhard Rost. I carried out necessary background search. The programming was performed by me with the help of Tobias Hamp. All calculations, data analyses and interpretations were done by me and Burkhard Rost. The manuscript was drafted by me, Tobias Hamp and Burkhard Rost.

2.2 Journal article. Goldberg T., Hamp T., Rost B. *Bioinformatics* 2012; 28(18):i458-i465

BIOINFORMATICS

Vol. 28 ECCB 2012, pages i458–i465
doi:10.1093/bioinformatics/bts390

LocTree2 predicts localization for all domains of life

Tatyana Goldberg[1,*,†], Tobias Hamp[1,†] and Burkhard Rost[1,2]

[1]TUM, Bioinformatik-I12, Informatik, Boltzmannstrasse 3, Garching 85748, Germany and [2]New York Consortium on Membrane Protein Structure (NYCOMPS) and Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

ABSTRACT

Motivation: Subcellular localization is one aspect of protein function. Despite advances in high-throughput imaging, localization maps remain incomplete. Several methods accurately predict localization, but many challenges remain to be tackled.

Results: In this study, we introduced a framework to predict localization in life's three domains, including globular and membrane proteins (3 classes for archaea; 6 for bacteria and 18 for eukaryota). The resulting method, LocTree2, works well even for protein fragments. It uses a hierarchical system of support vector machines that imitates the cascading mechanism of cellular sorting. The method reaches high levels of sustained performance (eukaryota: Q18=65%, bacteria: Q6=84%). LocTree2 also accurately distinguishes membrane and non-membrane proteins. In our hands, it compared favorably with top methods when tested on new data.

Availability: Online through PredictProtein (predictprotein.org); as standalone version at http://www.rostlab.org/services/loctree2.

Contact: localization@rostlab.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

1.1 Localization related to function

Archaea, bacteria and eukaryota form the three domains of life (Woese *et al*., 1990). Archaea and bacteria are prokaryotes, i.e. organisms that lack a nucleus and other membrane-bound organelles. Prokaryotic cells surround a single compartment by the plasma membrane (Gram-negative bacteria add an outer membrane). Eukaryotic cells are organized into several membrane-bound compartments. Subcellular localization is one aspect of cellular function as exemplified in the cellular component in the gene ontology (GO, Ashburner *et al*., 2000). Proteins contributing to the same physiological function often co-localize (Andrade *et al*., 1998; Jensen *et al*., 2002; Rost *et al*., 2003). Although proteins can be functional in different compartments (e.g. *importins* that shuttle other proteins into the nucleus), most proteins of known function complete their tasks as 'natives' of one particular compartment. For instance, many nuclear proteins are imported into the nucleus without being re-exported (Cokol *et al*., 2000); virulence-associated proteins are likely to be secreted in many bacterial pathogens (Durand *et al*., 2009). Increasing evidence suggests that proteins form temporary complexes to act in concert, resembling a macromolecular just-in-time production facility (Farhah Assaad TUM-WZW, personal communication). The knowledge of localization may, therefore, be important to understand protein interactions and cellular mechanisms.

1.2 Better annotations of function by predicting localization

The sequence-annotation gap refers to the gap between the number of proteins with known sequences and with comprehensive functional annotations. Next-generation sequencing explodes this gap despite increasing high-throughput experiments. Reliable automated predictions of protein function could counter this trend (Al-Shahib *et al*., 2007; Bairoch and Apweiler, 2000). Subcellular localization is one objective and easily definable aspect of function; many *in silico* prediction methods have been developed:

1. Sorting signals: Sorting signals (short motifs recognized by shuttle proteins) provide 'biologically meaningful' explanations for particular predictions. Most localization signals remain experimentally elusive (Nair and Rost, 2005) and many of the known signals have little coverage, i.e. allow the identification of very few proteins known to localize to that compartment (Wrzeszczynski and Rost, 2004). In addition, some proteins are sorted non-classically—not signal peptide triggered (Bendtsen *et al*., 2004).
2. Homology-based inference: The best localization predictions use annotations from close homologs (Nair and Rost, 2002b). This technique has limited reach because reliable inference requires high sequence similarity. It also has accuracy limitations: two 500-residue proteins may be sorted differently due to a 5-residue motif.
3. Text-based analyses: Text analysis-based methods infer localization from experimental information contained in the literature, such as PubMed abstracts (Brady and Shatkay, 2008) or from controlled vocabularies of curated databases, such as SWISS-PROT keywords (Nair and Rost, 2002a). All text-based methods are restricted in coverage as they rely on existing annotations.
4. *De novo: De novo* methods predict localization without requiring significant sequence similarity to annotated proteins. These methods are solely amino acid composition based (Chou, 2001; Park and Kanehisa, 2003; Reinhard and Hubbard, 1998).
5. Hybrid approaches combine several of these original four concepts (Blum *et al*., 2009; Briesemeister *et al*., 2009; Hoglund *et al*., 2006; Horton *et al*., 2007; Nair and Rost, 2005).

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Here, we present a novel sequence-based method for predicting the subcellular localization of all proteins in all domains of life. Our method addresses several shortcomings of existing approaches. (i) We provide a common framework for all domains of life and this framework is more robust with respect to sequencing mistakes than other methods. (ii) We increase the number of classes covered by a single consistent framework: 3 localization classes for archaea, 6 for bacteria and 18 for eukaryota. Predictions distinguish between integral trans-membrane and water-soluble globular (non-membrane) proteins. (iii) Similar to LocTree (Nair and Rost, 2005), we implemented a decision tree-like architecture of localization classes imitating the cellular protein sorting mechanisms. A tree-like structure accommodates the similarity of sorting signals specific to similar compartments (Alberts *et al*., 2007; Rusch and Kendall, 1995). (iv) We provide scores for the reliability of a prediction; these are crucial because they allow focusing on the most relevant results. All the above advantages were achieved without sacrificing performance. In our hands, LocTree2 performed significantly better than other methods on nearly all data sets tested.

2 METHODS

2.1 Data sets for development and evaluation

We extracted protein sequences with explicit annotations of subcellular localization from SWISS-PROT release 2011_04 (Bairoch and Apweiler, 2000). Excluded were annotations based on non-experimental findings (‘potential’, ‘probable’ or ‘by similarity’). Also excluded were proteins with multiple or ambiguous localization annotations (e.g. Gram-negative proteins annotated with ‘cell membrane’ could be in the inner or outer membrane). Proteins lacking the term ‘membrane’ were considered as ‘non-membrane’. Transmembrane proteins, i.e. proteins spanning the membrane at least once, were found using terms ‘single-pass’ or ‘multi-pass’. Through the NCBI taxonomy (Benson *et al*., 2010), proteins were assigned to one of the three domains (archaeal, bacterial or eukaryotic). Sequence bias was reduced through UniqueProt (Mika and Rost, 2003), applied independently for archaea, bacteria and eukaryota. This bias-reduction ascertained that no pair of proteins in the final set had BLAST2 (Altschul *et al*., 1990) E-value (EVAL) $\leqslant 10^{-3}$ or HSSP-value (HVAL) > 0 (Rost, 1999; Sander and Schneider, 1991). For alignments longer than 250 residues, HVAL < 0 implies that the maximal pairwise sequence identity was 20% (Rost, 1999). Filtering by HVAL and EVAL ensured that homology-based inference would be less accurate than our previous LocTree method (Nair and Rost, 2005). Alignments of fewer than 35 residues were removed, which is roughly the maximal length of known localization signals (Cokol *et al*., 2000). The final sets contained 59 archaeal, 479 bacterial and 1682 eukaryotic proteins (Supplementary Table S1).

2.2 Data sets for additional testing

After completing the development, we benchmarked our single best method against publicly available state-of-the-art methods. This involved the following independent test sets: (i) 28 bacterial and (ii) 52 eukaryotic proteins added to SWISS-PROT between releases 2011_04 and 2012_02; (iii) 43 *Arabidopsis thaliana* and (iv) 201 *Homo sapiens* proteins taken from LocDB (Rastogi and Rost, 2011). Proteins with HVAL > 5 to any previously used protein (including those discarded during the redundancy reduction) were excluded. This threshold corresponds to 25% pairwise sequence identity over 250 residues aligned. UniqueProt was used to reduce redundancy between the data sets and within each data set at HVAL > 0 and BLAST2 EVAL $\leqslant 10^{-3}$ with the minimum alignment length of 35 residues. We never used any of the remaining proteins (Supplementary Table S2) for any further improvement of our method. With the exception of LocTree, which used homology-based and text analysis-based predictions of SWISS-PROT proteins, and WoLF PSORT, which extracted an additional set of *Arabidopsis thaliana* proteins from Gene Ontology (Ashburner *et al*., 2000), the other methods tested here did not use any of the proteins in these independent test sets, as they were trained on data from SWISS-PROT releases before April 2011.

2.3 Additional data sets for comparison with LocTree

A question not addressed by the above data sets and comparisons is as follows: to which extent did our method benefit from the growth of the databases since 2005? In a separate analysis, all proteins for which localization had been annotated before 2005 served as training set and all from the above cross-validation set without sequence similarity (HVAL > 0 and EVAL $\leqslant 10^{-3}$) to this training set were used to compare LocTree2 and our previous method LocTree (Supplementary Table S3). No parameter optimization was applied when re-training our new method.

2.4 Prediction method

Each domain of life was considered as a separate learning problem yielding three different systems of decision trees (archaea: 3 classes, bacteria: 6 and eukaryota: 18; Fig. 1). Each leaf (rectangles) represents one localization class, and each internal node (circles) is a binary support vector machine (SVM). Most methodological aspects of the new method combine existing ideas. We briefly describe the main aspects here and leave the precise, formal definitions to the Supplementary Sections 1–3.

2.4.1 Input For each protein, sequence profiles were created by BLASTing (Altschul *et al*., 1997) queries against an 80% non-redundant database combining SWISS-PROT, TrEMBL (Bairoch and Apweiler, 2000) and the Protein Data Bank (Berman *et al*., 2000). Our method only used information available through these profiles.

2.4.2 Profile kernel Kernel methods (such as the SVM) differentiate between the input and the feature space. Here, the input space was spanned by all possible sequence-profile tuples. The feature space was implicitly given by the profile kernel (Kuang *et al*., 2004) that maps such a tuple to a vector indexed by all possible subsequences of length k from the alphabet of amino acids. Each element represents one particular k-mer and gives the number of identical k-mers with a score below a user-defined threshold σ. This score is calculated as the ungapped cumulative substitution score in the corresponding sequence profile. We can then define the profile kernel function as the dot product between the two k-mer vectors of the two sequence-profile tuples. Essentially, the method identifies stretches of k adjacent residues in the query that are most informative for the prediction of localization and then matches these in query protein.

2.4.3 SVM training SVMs were trained using a pre-computed kernel matrix of all training proteins. For the profile kernel, the matrix can be calculated very efficiently with the suffix tree-based ‘kernel trick’ introduced by the groups of Christina Leslie and Bill Noble (Leslie *et al*., 2004). We found other string kernels (Leslie *et al*., 2004; Lodhi *et al*., 2002) either slower in runtime or worse in performance (Supplementary Table S4). The SVM was implemented by the WEKA (Holmes *et al*., 1994) sequential minimal optimization (Platt, 1998). Platt Scaling (Platt, 1999) mapped the raw SVM score of the predicted class into a reliability between 0.5 and 1.0.

2.4.4 Tree-like hierarchy of SVMs The tree model (Fig. 1) was built by training binary SVM classifiers; each of those was trained on different sets of proteins. To this end, we first looked at one of the two child nodes of an internal node (e.g. internal node: root and child node: non-cytoplasmic; Fig. 1a) and collected all the training proteins of its leaf classes (e.g. EXT and PM; Fig. 1a). They were assigned to class A. Then we did the same for the second child node (e.g. CYT) and assigned its proteins to class B. Now, we could train the SVM of the parent node with the proteins in classes A and B. Repeating this for all internal nodes, we trained the entire tree model.

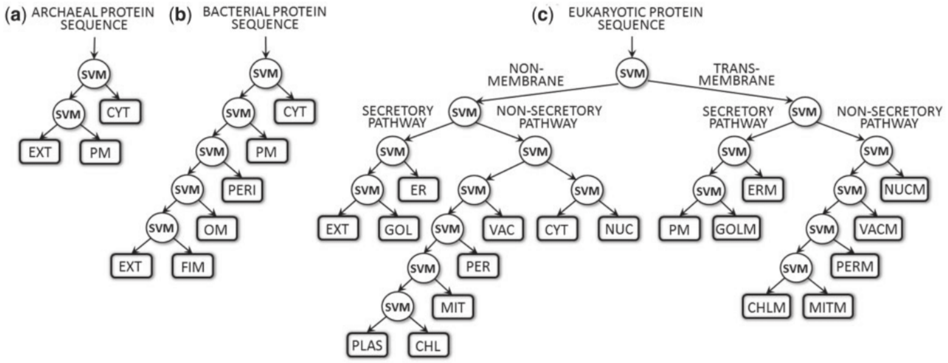


Fig. 1. Hierarchical architecture of LocTree2. The localization prediction follows a different tree for each of the three domains of life: **(a)** archaea, **(b)** bacteria and **(c)** eukaryota. Each hierarchy mimics the biological sorting mechanism in that domain (in eukaryotes membrane and non-membrane proteins are treated separately). The branches represent paths of the protein sorting, the leaves the final prediction of one localization class and the internal nodes are the decision points along the path. These decisions are implemented as binary support vector machines (SVMs). CHL, chloroplast; CHLM, chloroplast membrane; CYT, cytosol; ER, endoplasmic reticulum; ERM, endoplasmic reticulum membrane; EXT, extra-cellular; FIM, fimbrium; GOL, Golgi apparatus; GOLM, Golgi apparatus membrane; MIT, mitochondria; MITM, mitochondria membrane; NUC, nucleus; NUCM, nucleus membrane; OM, outer membrane; PERI, periplasmic space; PER, peroxisome; PERM, peroxisome membrane; PM, plasma membrane; PLAS, plastid; VAC, vacuole; VACM, vacuole membrane

2.4.5 Reliability index The reliability of the predicted class (leaf node) for a sequence-profile tuple was compiled as the product over the reliabilities of all parent nodes (as described in [Reinhardt and Hubbard, 1998]). We formed the LocTree2 reliability index (RI) by multiplying an integer of this value by 100. As the prediction confidence did not change for scores <20, the index was re-normalized accordingly.

2.5 Cross-validation

For training and testing, stratified 5-fold cross-validation was performed with each of the three sequence unique development data sets described before. This required several additional cross-validation layers to optimize various free SVM and multi-class learning parameters (Supplementary Section 1 for details). Note that we never used any information of the test split during a training phase. Entire rounds of cross-validation yielded comparisons to other multi-class learners (e.g. ENDs [Frank and Kramer, 2004]). Additionally, the influence of redundancy reduction was monitored; this suggested a controlled addition of redundancy after an initial reduction to be favorable (Supplementary Section 4).

2.6 Performance evaluation

Looking at predictions from the perspective of a single localization class L suggests various performance measures: the accuracy is the ratio between the number of correctly predicted proteins in localization L and all proteins predicted to be in L. Coverage is the ratio 'correctly predicted in L/all proteins observed in L'. Both values are combined in the geometric average gAv. The overall accuracy $Q(n)$ as the number of correctly predicted proteins across n classes divided by the number of observed proteins in these classes provides the perspective across all classes. Standard error for all measurements was estimated over 1000 bootstrap sets; i.e. randomly select n proteins without replacement from the original data set (in our experience, bootstrapping without replacement typically yields error estimates that are more conservative/long lived than those with replacement). For each bootstrapped set, the performance x_i is estimated (e.g. accuracy). These 1000 estimates provided the standard deviation of x_i with the typical standard error = standard deviation divided by $\sqrt{(n-1)}$, where n is the number of bootstrapped sets.

2.7 State-of-the-art prediction methods

We compared LocTree2 with the following publicly available state-of-the-art methods using default parameters.

CELLO 2.5 (Yu *et al*., 2006) is a system of SVMs that predicts localization of bacterial proteins to 5 classes and eukaryotic proteins to 12 classes. Predictions are based on sequence-derived features.

LocTree (Nair and Rost, 2005) predicts localization of non-membrane proteins from prokaryotes (three classes) and eukaryotes (six classes for plants and five for others) through the hierarchy of binary SVMs. The method uses features representing the entire protein and N-terminus specifics.

*MultiLoc*2 (Blum *et al*., 2009) uses SVMs that integrate sequence-based features with phylogenetic profiles and GO terms. It predicts 9 localization classes for animals/fungi and 10 plant classes (adding in chloroplast).

PolyPhobius (Kall *et al*., 2005) uses a hidden Markov model (HMM) for the prediction of transmembrane protein topology and signal peptides. It incorporates homology information for the increased prediction accuracy.

PSORTb 3.0 (Yu *et al*., 2010) predicts four classes for archaea/Gram-positives and five for Gram-negatives. It combines several classifiers by a Bayesian network to generate a final prediction of localization.

Scampi (Bernsel *et al*., 2008) predicts transmembrane protein topology through an HMM. Predictions are based on the experimental scale of position-specific amino acid contributions to the free energy of membrane insertion coupled with the positive-inside rule.

WoLF PSORT (Horton *et al*., 2007) is a *k*-nearest neighbor classifier that predicts 12 localization classes for eukaryotes from sequence-based features. Similar to its predecessors (PSORT), it uses a tree

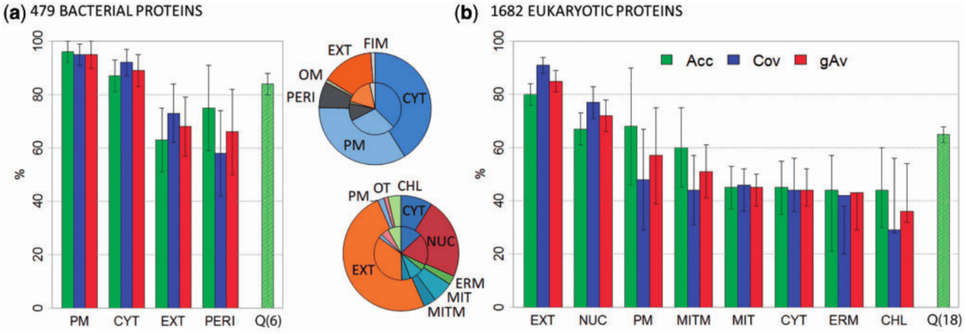


Fig. 2. High performance in cross-validation. For the cross-validation sets (**a**: averages over 479 bacterial proteins and **b**: averages over 1682 eukaryotic proteins), LocTree2 reached high levels of sustained performance. Overall, performance tended to correlate with the number of representatives (pie charts: inner ring: composition in the corresponding data set and outer ring: composition in correct predictions). Exceptions were membrane bound classes in eukaryotes for which the performance tended to be better than that for the corresponding non-membrane bound class (e.g. MIT = mitochondrial proteins versus MITM = membrane-linked mitochondrial proteins). Localization classes as in Figure 1; performance measures: Acc, accuracy; Cov, coverage; gAv, geometric coverage of Acc and Cov; *Q*, overall prediction accuracy (Q6 for six and Q18 for 18 classes). Standard errors were estimated by bootstrapping (see Section 2). Classes with less than 20 members were excluded

hierarchy resembling cellular sorting and a battery of established prediction methods.

Unlike all others, *PolyPhobius* and *Scampi* do not aim at predicting localization. Instead, they focus on the prediction of which residues are inserted as transmembrane helices into the lipid bilayer. In the context herein, those two methods are compared to demonstrate that LocTree2 could even stand up to specialists that optimize the distinction of membrane and non-membrane proteins in their own domain of specialization.

3 RESULTS AND DISCUSSION

3.1 Three prediction trees for three domains of life

Our first hierarchal method, LocTree (Nair and Rost, 2005), used a concept initially introduced by the work on PSORT carried by Paul Horton and initiated by Kenta Nakai and Minoru Kanehisa (Horton *et al*., 2007; Nakai and Horton, 1999; Nakai and Kanehisa, 1991). For LocTree2, many alternative trees were tested. Trees mimicking the cellular protein trafficking using binary models at the internal nodes (Fig. 1) were similar in performance but much faster than other multi-class schemes, for example ENDs (Frank and Kramer, 2004) (Supplementary Table S5). Starting at the root classifier (e.g. non-membrane/trans-membrane; Fig. 1c), the decisions at each node are followed until reaching a leaf (e.g. mitochondria membrane [MITM]). This leaf corresponds to the predicted localization class (development set in Supplementary Table S1).

3.2 Cross-validated Q18 = 65% for eukaryotes

The first decision for eukaryotic proteins was: does it have an integral transmembrane region or not (Fig. 1c). This decision was correct for over 90% of all proteins (Supplementary Figure S1b). Both membrane and non-membrane proteins were further classified into ‘secreted’ and ‘not secreted’; this decision reached Q4 = 83% accuracy (Q4 = four state accuracy, see Section 2 for definition of *Qn*; Supplementary Figure S1b). Descending the tree toward the leaves that represent the final predictions, the distinction between intra-cellular and secretory pathway into 10 classes for non-membrane and 8 classes for transmembrane proteins was less accurate (Q8 = 75%; Supplementary Figure S1b). The class with most observations (extra-cellular: 35% of data) was also predicted best (accuracy: 80%, coverage: 91%, Fig. 2b, Supplementary Table S6) followed by nuclear proteins (accuracy: 67%, coverage: 72%). The overall accuracy for 18 classes Q18 reached 65% (18-state accuracy, Fig. 2b).

Overall, performance correlated with the amount of available experimental information (Fig. 2b: inner and outer pies very similar), with the important exception that membrane-bound proteins tended to be predicted more accurately than their corresponding non-membrane bound neighbors (e.g. mitochondria [MIT] versus MITM in Fig. 2b).

3.3 Highest numerical performance for prokaryotes

LocTree2 performed very well in the cross-validation of archaea (three classes) with overall levels of accuracy and coverage numerically suggested to reach 100% (Supplementary Table S7). These numbers most likely over-estimate performance due to the limited data. For bacteria (six classes), the overall accuracy was 84% (Fig. 2a); the most accurate sub-classification was the sorting into plasma membrane (accuracy: 96%, Fig. 2a, Supplementary Table S6) followed by cytosol (accuracy: 87%).

3.4 Performance best for more reliably predicted proteins

One way to focus on more reliable predictions is to compile a consensus for alternative methods. Often, method internal reliability indices are far superior at spotting the best predictions than combinations of different methods (Eyrich *et al*., 2003). LocTree2 computed the reliability index (RI) as the joint probability over all individual SVM scores (see Section 2, Fig. 3). For instance, the

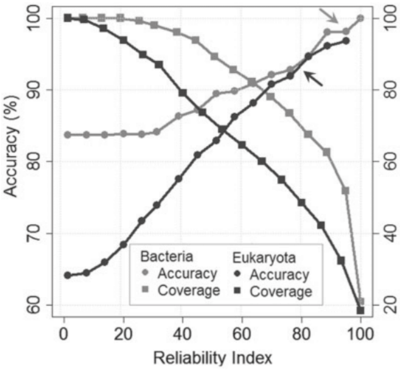


Fig. 3. More reliable predictions better. The curves show the percentage accuracy/coverage for LocTree2 predictions above a given threshold in the reliability index (from 0 = unreliable to 100 = most reliable). True positives are the number of correct predictions with reliability indices above the given threshold, false negatives are the number of correct predictions with reliability indices below the threshold and false positives are the number of wrong predictions with reliability indices above the threshold. The curves were obtained on cross-validated test sets of bacterial (gray line) and eukaryotic (black line) proteins. Half of all eukaryotic proteins are predicted at RI>80; for these, Q18 is above 92% (black arrow). As the number of localization classes is lower for bacteria, the corresponding number in accuracy is higher (Q6 is above 95% at 50% coverage, gray arrow)

50% of the proteins with highest reliability reached levels of overall accuracy Q6 = 98% for bacteria (Fig. 3, gray arrow) and Q18 = 92% for eukaryota (Fig. 3, black arrow). To pick another point, almost 40% of all eukaryotic proteins were predicted at RI > 85; for these, Q18 was above 95%. Thus, two in the top 40 predictions in 100 were wrong in one of 18 states (e.g. nuclear instead of nuclear membrane).

3.5 LocTree2 competitive for new proteins

There is no value in comparing LocTree2 with other methods based on values for performance published because of the differences in, for example data sets and cross-validation setups. Comparisons based on running other methods on our data are also problematic due to possible overlap in training and due to possible performance over-estimates of our own method. The only meaningful way is to use proteins that are non-redundant with respect to each other and with respect to any protein used for the development of the methods tested. Toward this end, we collected the most recently added annotations in SWISS-PROT. The price for this ‘clean’ comparison was the tiny data set: 28 bacterial and 52 eukaryotic proteins after redundancy reduction (explaining high standard errors in Table 1).

CELLO 2.5 and PSORTb 3.0 classified bacterial proteins into five classes and LocTree into three. This was accounted for by grouping bacterial extra-cellular and fimbrium proteins into one common class for predictions using these external methods. We separated Gram-positive from Gram-negative bacterial proteins according to Yu *et al*. (2010) for a comparison with PSORTb 3.0.

Eukaryotic proteins were classified into twelve classes for CELLO 2.5 and WoLF PSORT, into ten classes for MultiLoc2 and into six classes for LocTree. We excluded vacuolar proteins for MultiLoc2 and plasma membrane proteins for LocTree (thereby providing over-optimistic upper performance levels for those methods). WoLF PSORT may predict multiple localizations, and we always took the right one for performance estimates (it was verified that this did not impact estimates significantly). WoLF PSORT and CELLO 2.5 distinguish cytoskeleton and cytoplasm; here, both were considered as cytoplasmic. Another issue was that other methods do not distinguish membrane from non-membrane proteins. Thus, we merged these two classes, i.e. treated nuclear and nuclear-membrane proteins identically, although this approach implicitly sacrificed one of the important strengths of our new method, namely the distinction of these.

The ‘New SWISS-PROT’ bacterial and eukaryotic sets were too small to clearly identify the top performing method given the standard error. However, LocTree2 compared favorably to other state-of-the-art methods (Table 1). Performance estimates for the newly annotated proteins tended to be lower than the values published (except for LocTree and MultiLoc2). For LocTree2, the overall accuracy was similar for the cross-validation experiment (84% ± 4% for bacteria and 65% ± 3% for eukaryota; Fig. 2, Supplementary Table S6) and for the new proteins (86% ± 16% for bacteria and 65% ± 14% for eukaryota; Table 1).

3.6 LocTree2 would already have performed well in 2005

Another way to compare two prediction methods is to train and test on the same data set. We trained a version of LocTree2 on proteins for which localization was known when LocTree was trained and tested both on proteins from our newer cross-validation set without sequence similarity to the training set (see Section 2 and Supplementary Table S3). LocTree2 outperformed LocTree reaching levels of overall accuracy Q3 = 80% ± 13% for bacteria and Q6 = 61% ± 8% for eukaryota (LocTree: Q3 = 62% ± 18% and Q6 = 54% ± 8%). Thus, the improvement of LocTree2 originated mainly from the underlying method advancement. LocTree2 trained on the 2011 data reached Q6 = 62% ± 8% and Q18 = 60% ± 9% which is within the standard error of what was obtained on the full cross-validation set (Supplementary Table S6).

3.7 High-throughput data ambiguous?

LocDB collects localization annotations mostly from high-throughput experiments; it provided two data sets for the comparison of methods: one for the plant *Arabidopsis thaliana* and the other for *Homo sapiens*. Both sets were redundancy reduced, with respect to each other and with respect to SWISS-PROT version 2011_04. For all the LocDB proteins, all methods appeared to perform substantially worse than for the already ‘tough’ set of newly annotated SWISS-PROT proteins. For the plant, LocTree2 out-performed others (Table 1). Not so for human: WoLF PSORT reached Q8 = 45% ± 8% (versus LocTree2 Q8 = 42% ± 8%). One-third of the correct predictions from WoLF PSORT were for cytoplasmic proteins, which was overall, the most populated class for human proteins in LocDB (Supplementary Table S2).

How to interpret the data from LocDB? As most annotations in LocDB originate from high-throughput experiments, it is very

Table 1. Performance comparison on independent data sets

Method	New SWISS-PROT					LocDB					
	Bacteria (28)		*Eukaryota (52)*			*A. thaliana (43)*			*H. sapiens (201)*		
	Q(5)	Q(3)	Q(9)	Q(6)	Q(5)	Q(9)	Q(8)	Q(6)	Q(8)	Q(7)	Q(6)
LocTree2	**86 ± 16**	**86 ± 18**	**65 ± 14**	**66 ± 16**	**66 ± 15**	**37 ± 18**	**44 ± 21**	**49 ± 20**	42 ± 8	**44 ± 9**	**51 ± 9**
CELLO v. 2.5	57 ± 22	—	46 ± 16	—	—	26 ± 18	—	—	40 ± 8	—	—
WoLF PSORT	—	—	62 ± 14	—	—	19 ± 15	—	—	**45 ± 8**	—	—
PSORTb 3.0	71 ± 21	—	—	—	—	—	—	—	—	—	—
MultiLoc2	—	—	—	60 ± 16	—	—	24 ± 18	—	—	42 ± 9	—
LocTree		77 ± 21	—	—	62 ± 17	—	—	24 ± 18	—	—	48 ± 9

Data 'New SWISS-PROT': 28 sequence-unique bacterial and 52 eukaryotic proteins added to SWISS-PROT between releases 2011_04 and 2012_02 (sequence uniqueness was ascertained both within this set and from any protein in this set to any other protein previously in SWISS-PROT). *Data 'A. thaliana' and 'H. sapiens'*: 43 *Arabidopsis thaliana* and 201 *Homo sapiens* proteins from the LocDB database (as for 'New SWISS-PROT': sequence unique with respect to itself and to SWISS-PROT 2011_04). *Qn*, the overall prediction accuracy in *n* classes; highest value in each column in bold; values ± standard error (see Section 2).

likely that LocDB contains proportionally more errors than SWISS-PROT. All methods by far outperformed random, implying that for random annotation mistakes they would appear to be mostly wrong. Thus, the higher error rate in LocDB might explain why all methods perform worse for the LocDB than for the SWISS-PROT data. Put differently, for a task with over six classes and the given number of proteins, a few mistakes can reduce the average considerably. On the other hand, we might also suspect that high-throughput experiments discover a reality invisible to traditional experimental methods and some of those invisible facts might reveal new sorting mechanisms. Such hidden mechanisms might or might not be 'discovered' by prediction methods. If not, those would explain many incorrect predictions. Supposedly, most experts would be very surprised if the second argument (new mechanism) dominated over the first (annotation mistakes of high-throughput experiments). Most likely there is a little bit of both, but we have no means of gauging the relative proportions. Zooming into annotations with several evidences brought the numbers closer, i.e. 'increased' the performance, but this was achieved at raising the standard errors to meaningless values (Supplementary Table S8).

We illustrate the situation for a few extreme predictions. (i) 'Transmembrane emp24 domain-containing protein 3' (SWISS-PROT TMED3_HUMAN) is annotated as Golgi apparatus by LocDB; LocTree2 maps it to the endoplasmic reticulum (ER) membrane with extremely high reliability (RI=99). This protein belongs to a family of p24 membrane proteins localizing to the ER and to the Golgi complex (Jenne *et al*., 2002). Thus, both LocTree2 and LocDB annotations are correct. (ii) 'Protein canopy homolog 2' (CNPY2_HUMAN) is annotated as cytoplasmic in LocDB; LocTree2 predicts ER (RI=73). We found experimental evidence for localization to the ER in HeLa cells (Hirate and Okamoto, 2006). In this case, LocTree2 is correct and LocDB is not. (iii) 'Methylosome subunit pICln' (ICLN_HUMAN) is classified as plasma membrane in LocDB, whereas LocTree2 predicts nuclear (RI=55). We could not find any additional information for this case in PubMed, but the protein localization annotation in SWISS-PROT is nuclear. (iv) 'COMM domain-containing protein 1' (COMD1_HUMAN) is classified as secreted in LocDB, whereas LocTree2 predicts nuclear (RI=50). Again, closer inspection revealed experimental evidence for this protein to be nuclear (Burstein *et al*., 2005).

It remained unclear what to conclude from the above examples. The predictions judged as incorrect by LocDB but having very high reliability scores indicate that the low performance inverts the real picture: rather the annotations are wrong or ambiguous than the strong predictions. For a set of weakest predictions, we observed the opposite. For example (i) 'Stress-associated endoplasmic reticulum protein 1' (SERP1_HUMAN) is annotated as ER correctly in LocDB, but LocTree2 maps it to mitochondria with very low reliability (RI = 6). (ii) 'Spermatogenesis-associated protein 19, mitochondrial' (SPT19_HUMAN) is classified as mitochondrial correctly in LocDB again, whereas LocTree2 predicts nuclear (RI = 13). A more detailed analysis might succeed in quantifying to which extent the consistent drop in performance for the LocDB data sets reveals more about problems of high-throughput experiments than of *mega-throughput* computations.

3.8 Accurate distinction between membrane and non-membrane

As reported before, the SVM that distinguishes between non-/trans-membrane proteins in eukaryotes achieved an overall accuracy of 94% ± 2% (Supplementary Figure S1b). This performance was similar to what PolyPhobius achieved on the same data set (95% ± 1%). PolyPhobius appears to be the best expert method that targets the prediction of integral membrane helices directly (Kloppman E., Reeb J. and Rost B., unpublished data). LocTree2 correctly classified all plasma membrane proteins from archaea (Supplementary Table S7), but the data set was too small to provide meaningful performance estimates. For bacterial proteins, the plasma membrane/non-membrane distinction reached 96% ± 4% accuracy (Fig. 2a, Supplementary Table S6). Scampi, the most accurate method for predicting trans-membrane proteins in prokaryotes (Kloppman E., Reeb J. and Rost B., unpublished) was significantly less accurate (89% ± 3%) for the same data.

3.9 Advantage over existing methods for sequencing errors

All prediction methods were also benchmarked on protein fragments as they may result from erroneous assembly or wrong gene predictions common in genome projects (Brent and Guigo, 2004). The latter being a special problem for the detection of N-terminal signals because of the wrong predictions of gene starts common when using gene prediction software. Three different ‘models’ simulated worst-case scenarios (over-estimating sequencing mistakes): cleaving off (i) 30 N-terminal residues for all proteins, (ii) 30 C-terminal residues and (iii) randomly picking positions to cleave one third of the sequence. The least ‘damage’ was done for the C-term cleavage with LocTree2’s accuracy dropping to 60% ± 2% (Supplementary Table S9), which was still within the standard error of what was obtained using the full-length sequences. For other prediction methods, performance dropped much more. Our method also significantly outperformed its competitors on the N-term cleaved sequences and on the sequences with randomly cleaved fragments, reaching the levels above 53% ± 2% accuracy (Supplementary Table S9). This is still accurate enough to provide reliable first estimates of localization for genomic sequences.

4 CONCLUSION

The method introduced here, LocTree2, predicts protein subcellular localization through a consistent new framework that ignores many of the relevant features needed for the success of previous methods (such as no predicted aspects of protein structure and function). Nevertheless, it seemed to reach high levels of sustained performance aside from adding new aspects. Among the novel aspects was the large number of 18 localization classes predicted for eukaryota, 6 for bacteria and 3 for archaea. LocTree2 outperformed other methods on almost all data sets tested, implicating an improved ability to capture localization signals in the protein sequence. One example for the success in plucking implicit information is the high precision in the distinction between membrane and globular water-soluble proteins. Our implicit distinction appeared as good as that of the best expert method for predicting integral membrane helices. Another important novelty is the robustness of the method against sequencing errors and its success when applied to protein fragments. This is particularly important in light of high-throughput sequencing, of analyzing ancient DNA with short reads and of the fact that almost 80% of all proteins have multiple domains. This power along with the overall improvement in performance may recommend this new tool as an ideal starting point for comparing the proteomes between organisms and for using localization predictions to aid the prediction of protein function. We imagine that the framework for the method will prove extendable and that future methods will become better simply by using more experimental data and more sequences.

ACKNOWLEDGEMENTS

Thanks to Laszlo Kajan (TUM), Guy Yachdav (Columbia) and Tim Karl (TUM) for initial help with LocTree and for help with the computer cluster; to Marlena Drabik (TUM) for administrative support. Particular thanks to Shruti Rastogi (Columbia) for providing the LocDB data and Yana Bromberg (Rutgers) for help with the manuscript. We further thank Sebastian Briesemeister (Univ Tuebingen) for kindly providing MultiLoc2 predictions. Last, not least, thanks to Rolf Apweiler (UniProt, EBI), Amos Bairoch (CALIPHO, SIB), Ioannis Xenarios (UniProt, SIB) and their crews for maintaining excellent databases and to all experimentalists who enable analysis by making their data publicly available.

Funding: This work was supported by a grant from the Alexander von Humboldt foundation through the German Federal Ministry for Education and Research (BMBF).

Conflict of Interest: None declared.

REFERENCES

Al-Shahib,A. *et al.* (2007) Predicting protein function by machine learning on amino acid sequences—a critical evaluation. *BMC Genomics*, **8**, 78.

Alberts,B. *et al.* (2007) *Molecular Biology of the Cell.* Garland Science, New York.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andrade,M.A. *et al.* (1998) Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.*, **276**, 517–525.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Bendtsen,J.D. *et al.* (2004) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Design Select.*, **17**, 349–356.

Benson,D.A. *et al.* (2010) GenBank. *Nucleic Acids Res.*, **38**, D46–D51.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bernsel,A. *et al.* (2008) Prediction of membrane-protein topology from first principles. *Proc. Natl. Acad. Sci. USA.*, **105**, 7177–7181.

Blum,T. *et al.* (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, **10**, 274.

Brady,S. and Shatkay,H. (2008) EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pacific Symp. Biocomput.*, **2008**, 604–615.

Brent,M.R. and Guigo,R. (2004) Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.*, **14**, 264–272.

Briesemeister,S. *et al.* (2009) SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J. Proteome Res.*, **8**, 5363–5366.

Burstein,E. *et al.* (2005) COMMD proteins, a novel family of structural and functional homologs of MURR1. *J. Biol. Chem.*, **280**, 22222–22232.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**, 246–255.

Cokol,M. *et al.* (2000) Finding nuclear localization signals. *EMBO Rep.*, **1**, 411–415.

Durand,E. *et al.* (2009) Structural biology of bacterial secretion systems in gram-negative pathogens—potential for new drug targets. *Infect. Disord. Drug Targets*, **9**, 518–547.

Eyrich,V.A. *et al.* (2003) CAFASP3 in the spotlight of EVA. *Proteins Struct. Funct. Bioinformatics*, **53**(Suppl 6), 548–560.

Frank,E. and Kramer,S. (2004) Ensembles of nested dichotomies for multi-class problems. In *ICML-2004*, pp. 305–312, ACM Press.

Hirate,Y. and Okamoto,H. (2006) Canopy1, a novel regulator of FGF signaling around the midbrain-hindbrain boundary in zebrafish. *Curr. Biol.*, **16**, 421–427.

Hoglund,A. *et al.* (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165.

Holmes,G. *et al.* (1994) WEKA: A Machine Learning Workbench. *Proceedings of Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, pp. 357–361.

Horton,P. *et al.* (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.

Jenne,N. *et al.* (2002) Oligomeric state and stoichiometry of p24 proteins in the early secretory pathway. *J. Biol. Chem.*, **277**, 46504–46511.

Jensen,L.J. *et al.* (2002) Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, **319**, 1257–1265.

Kall,L. *et al.* (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21**(Suppl 1), i251–i257.

Kuang,R. *et al.* (2004) Profile-based string kernels for remote homology detection and motif extraction. *Proceedings/IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference*, pp. 152–160.

Leslie,C.S. *et al.* (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Lodhi,H. *et al.* (2002) Text classification using string kernels. *J. Machine Learn. Res.*, **2**, 419–444.

Mika,S. and Rost,B. (2003) UniqueProt: Creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Nair,R. and Rost,B. (2002a) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18**(Suppl 1), S78–S86.

Nair,R. and Rost,B. (2002b) Sequence conserved for subcellular localization. *Protein Sci. Publ. Protein Soc.*, **11**, 2836–2847.

Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85–100.

Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.

Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins*, **11**, 95–110.

Park,K.J. and Kanehisa,M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.

Platt,J.C. (1998) Fast training of support vector machines using sequential minimal optimization. In Schölkopf,B. Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA, pp. 185–208.

Platt,J.C. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likehood methods. In Peter,J. Bartlett, *et al*. (eds.) *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, pp. 61–74.

Rastogi,S. and Rost,B. (2011) LocDB: experimental annotations of localization for *Homo sapiens* and *Arabidopsis thaliana*. *Nucleic acids Res.*, **39**, D230–D234.

Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic acids Res.*, **26**, 2230–2236.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Rost,B. *et al.* (2003) Automatic prediction of protein function. *Cell Mol. Life Sci.*, **60**, 2637–2650.

Rost,B. *et al.* (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–W326.

Rusch,S.L. and Kendall,D.A. (1995) Protein transport via amino-terminal targeting sequences: common themes in diverse systems. *Mol. Membr. Biol.*, **12**, 295–307.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Woese,C.R. *et al*. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA.*, **87**, 4576–4579.

Wrzeszczynski,K.O. and Rost,B. (2004) Annotating proteins from endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes. *Cell. Mol. Life. Sci.*, **61**, 1341–1353.

Yu,C.S. *et al*. (2006) Prediction of protein subcellular localization. *Proteins*, **64**, 643–651.

Yu,N.Y. *et al*. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.

Supporting online material for: LocTree2 predicts localization for all domains of life

Tatyana Goldberg, Tobias Hamp & Burkhard Rost

SOM Section 1: LocTree2 development workflow

We extracted sets of archaeal, bacterial and eukaryotic proteins together with their experimentally determined annotations of subcellular localization from the SWISS-PROT database (Bairoch and Apweiler, 2000) (Methods). We internally homology reduced these data sets to avoid homology-based inference of subcellular localization (Methods). We built three separate classification models, one for each domain of life (Fig. 1). This means we carried out the entire following procedure three times.
As a first step, we divided a data set into five equally sized subsets (Methods) in order to train and test various classification models via cross-validation. Applying "stratification", we made sure that classes had about the same size in each of the five sets. There was nothing special in the way we carried out the cross-validation: in one fold, four subsets were used for training and one for testing. Then, all subsets were rotated such that each subset was used for testing exactly once. Finally, we averaged the performances over all test sets. We always introduced a certain degree of homology within each of the training sets in order to increase them in size (SOM Section 4). This was found to be beneficial and did not compromise the similarity between training and test proteins.

In this work, we essentially wanted to study the power of support vector machines (SVMs) (Cortes and Vapnik, 1995) in combination with string kernels for subcellular localization prediction. A kernel is the most crucial part of a SVM: it determines the 'feature space', i.e. the space into which the objects under consideration (here: protein sequences) are mapped and where they are linearly separated. The better the kernel, the better we will be able to discriminate between proteins from different localization classes. As we have to apply a kernel function during the developmental phase many times for training and testing, its speed also majorly determines the degree with which we can optimize, e.g. free parameters and how fast new, unseen proteins can be classified.
In our case, the problem of optimizing free parameters was ubiquitous: it started with the choice of the multi-class classification scheme and the underlying kernel function. Each kernel, in turn, had at least two other parameters. Additionally, we had to optimize the SVM complexity parameter C and perform Platt-Scaling (Platt, 1999) for each binary SVM (Platt Scaling converts SVM scores into probabilities). Because optimizing all these parameters at once would have created impossible amounts of value combinations and learning tasks, we decided to solve the problem in three major steps. In other words, given parameters A,B,C,D this essentially meant that we first optimized parameters A,B, and then C,D with the best values for A,B. We understand that this does not find a 'global optimum' in the sense of the best parameter combination. As we later show, however, the 'local optimum' that we find is better than most, if not all, other current subcellular localization predictors. Besides, we always put great emphasis on not allowing test data to 'leak' into the training data, what sometimes led to quite complicated setups. The WEKA package (Holmes, et al., 1994) was an invaluable help in this process and we provide command line calls of each step upon request.

In our first step, we compared the performances of three different string-based kernel functions, namely the String Subsequence Kernel (Lodhi, et al., 2002), the Mismatch Kernel (Leslie, et al., 2004) and the Profile Kernel (Kuang, et al., 2004; SOM Section 2). At this early stage of our evaluation, we decided to only use the simplest and fastest multi-class classification approach one-against-all (Allwein and Singer, 2000). We performed one entire cross-validation (above) for each kernel and then chose the winner for subsequent steps. In each cross-validation fold, we optimized the different kernel-specific parameters.

We optimized kernel specific parameters in a nested ten-fold cross-validation: an 'outer' training set (comprising 4/5 of the original data set; before) was again split into 10 'inner' partitions. We then trained a one-against-all model with a particular parameter combination on 9 of these splits and tested it on the remaining one. This was repeated ten times by rotating through the 'inner' folds and we obtained the average performance for this parameter-combination and 'outer' training set. Then, we changed the parameter combination and repeated the nested cross-validation. Finally having calculated the performance for all parameter combinations, we chose the one leading to the best performance. This combination was then used to train a model using all 10 inner splits and to predict proteins in the outer test split.
We repeated this procedure five times by rotating through the outer cross-validation folds in order to predict all proteins. As mentioned before, we chose the kernel leading to the highest overall accuracy as the winner. This was the Profile Kernel. Depending on the kingdom (archaea, bacteria or eukaryota), it was found to be either within the standard errors of other kernel functions or significantly more accurate. Additionally, it was much faster in runtime compared to other kernel functions (data not shown).

Having found the best kernel in this way, we assessed in our second step the performance of different multi-class classification approaches. These were One-Against-All (Allwein and Singer, 2000), Ensembles of Nested Dichotomies (Frank and Kramer, 2004), Ensembles of Class Balanced Dichotomies (Dong, et al., 2005), Ensembles of Data Balanced Dichotomies (Dong, et al., 2005) and Nested Dichotomies of a Fixed Structure (Methods; Fig. 1). We briefly introduce them in SOM Section 3. We again evaluated the performance for each of the five splits of the training set in a stratified 10-fold cross-validation. The only difference to before was that instead of three different kernels and one multi-class approach, we now evaluated one kernel (the Profile Kernel) and five multi-class classification schemes.
We observed a superiority of dichotomies-based classification approaches over one-against-all, but could not find a significant difference in prediction performance among them. However, measuring the classification speed (the number of protein sequences processed per minute) revealed a significant advantage of nested dichotomies with a fixed structure (Fig. 1; Table SOM_3) and we chose these models for our final step (note that each dichotomy is different for the three kingdoms because they have different localization classes).

So far, we have found the optimal kernel (the Profile Kernel) and the multi-class classification scheme (Nested Dichotomies of a Fixed Structure). We have still neglected the optimization of the SVM cost parameter C and have not used Platt Scaling. In our third and last step, we included them in the form of two additional 5-fold cross-validation layers for each binary SVM (increasing the overall number of layers to 4). For example, in the second outer cross-validation fold, the seventh inner fold, the Profile Kernel parameter combination (k=2 and σ=6), and the root node of the nested dichotomy of a fixed structure for archaea, we have a binary data set for the corresponding SVM (cytoplasmic vs. non-cytoplasmic). In order to find its best parameter C, we performed a 5-fold cross-validation for each possible value (0.01, 0.1, 1.0, 10, 100, 1000). For each of these 5*6=30 folds, we performed Platt-Scaling, using a 5-fold cross-validation again.

In Fig. 2, we present the average performance over the five outer folds of our model after this last step.

The final models (one for each kingdom), which we evaluated against current state-of-the-art prediction methods for subcellular localization (Methods), and that we installed on our server, were obtained after a final re-training using all of the five outer folds as training data. The Profile Kernel parameters for our final models were: k=3 and σ=5 for archaea, k=5 and σ=9 for bacteria, k=6 and σ=11 for eukaryota.

An in-depth analysis of these models with respect to unknown signal peptides is theoretically possible, but necessary methodologies are yet to be developed. Nevertheless, we provide a preliminary analysis of one of our SVMs in SOM Section 5.

SOM Section 2: The Profile Kernel.

There are a number of sequence-based kernel functions designed for protein classification tasks. In this work, we applied and compared the String Subsequence Kernel (Lodhi, et al., 2002), the Mismatch Kernel (Leslie, et al., 2004) and the Profile Kernel (Kuang, et al., 2004). The main idea behind them is to compare two protein sequences by looking at the number of common subsequences of a fixed length. No biological knowledge is incorporated, in the sense that protein sequences are simply represented as strings of amino acids. In the following, we introduce the main principles behind the Profile Kernel, the kernel function selected to be used for the LocTree2 classification system.

Evolutionary sequence profile. The key feature of the Profile Kernel, as the name already states, is the use of protein sequence profiles. Such a profile is estimated by aligning a target sequence against a group of homologous (similar) sequences, e.g. obtained via BLAST (Altschul, et al., 1990), and compiling the conservation of each amino acid at each alignment position into a score. Many different ways have been proposed on how to compute these scores, with the most popular variant arguably being the one by Altschul et al. and implemented in PSI-BLAST (Altschul et al., 1997). In the profile kernel, the score is simply the negative logarithm of the amino acid frequency at a particular position, slightly 'smoothed out' by pseudo counts (pseudo amino acid probabilities estimated from the training data [Kuang, et al., 2004]). Consequently, in the *n*20* scoring matrix (*n* being the length of the target sequence and *20* the size of the amino acid alphabet), a value around 0.0 indicates that the respective amino acid has often been observed at a particular position, whereas higher values mean the opposite, i.e. weak or no conservation. We obtained position specific frequency matrices from PSI-BLAST by querying the target sequence against a redundancy reduced combination of SWISS-PROT, TrEMBL (Bairoch and Apweiler, 2000) and PDB (Berman, et al., 2000) (Methods).

Computation of the Profile Kernel. The Profile Kernel makes use of an evolutionary profile P_s of a sequence s**.** The user has to define the length of the subsequences to consider (k) and the conservation threshold *σ.* The latter defines a filter for *k*-mers which exhibit high sequence diversity.
More formally: Given *k*-mer $m(s, j)$,

$$m(s, j) \; = s\big[j+1 : j+k\big] = s_{j+1} \ldots s_{j+k} \text{ with } \; 0 \le j \le |s| - k$$

The kernel looks at the corresponding part of the profile ($P_{(s,j)}$, *i.e.* the profile P_s reduced to substitution scores between residues j+1 and j+k) and determines all *k*-mers with a cumulative substitution score below *σ* :

$$S_j = \left\{ x \;\middle|\; x \in \Sigma^k \wedge -\sum_{i=1}^{k} \log P_{(s,j)}(i, x_i) < \sigma \right\}$$

with Σ being the alphabet of 20 amino acids and $P_{(s,j)}(i, x_i)$ the frequency of amino acid x_i at position i in the sub-profile $P_{(s,j)}$.

Then, we can define the feature map of the kernel:

$$\Phi^{x}(P_{s}) = \sum_{j=0}^{|s|-k} I(x \in S_{j})$$

with $x \in \Sigma^{k}$,

$\Phi^{x}(P_{s})$ indicating the value of *k*-mer x in the feature vector $\Phi(P_{s})$ and

$I(x) = 1$ if $x \in S_{j}$.

Note that $\Phi(P_{s})$ has $|\Sigma|^{k} = 20^{k}$ dimensions. Consequently, the Profile Kernel is defined as the dot product of two feature vectors:

$$k(P_{s1}, P_{s2}) = \Phi(P_{s1}) \cdot \Phi(P_{s2})$$

Directly following the procedure above when implementing the kernel would quickly result in unfeasible runtime and memory requirements. Luckily, we can entirely avoid explicitly mapping a profile into the *k*-mer feature space and directly compute the kernel. This is commonly known as the ‘kernel trick’. Additionally, we can combine the computation of kernel values and create the entire kernel matrix in one operation. The kernel matrix is an all-against-all comparison of each protein in the data set. Each cell contains the kernel value of the two proteins under comparison. SVMs either create it on their own during training or receive it from the user.

The Profile Kernel applies an efficient data structure that is built on all *k*-mers, called suffix trie, for the efficient computation of the kernel matrix. *k*-long profiles are stored on the path from the root to the leaf. An internal node of depth *d* stores a set of pointers to all *k*-length profiles $P_{(s,j)}$, whose current cumulative conservation scores are less than the *σ* threshold.

More specifically:

$$n_{p} = \left\{ P_{(s,j)} \;\middle|\; -\sum_{i=1}^{d} \log P_{(s,j)}(i, s_{i}) < \sigma \wedge m(s,j)[1:d] = seq(n) \right\}$$

where n_{p} is the set of k-long sub-profiles stored at node n and $seq(n)$ is the sequence induced by the path from the root to node n.

Only processing the profiles remaining at the leaf nodes, we can save a lot of computation and compute many kernel values at the same time. We refer to Leslie et al, 2004 and Kuang et al. 2005 for a more detailed description of the procedure.

Generally, the complexity of computing a Profile Kernel value $\mathrm{k}(\mathrm{P}_{\mathrm{s1}}, \mathrm{P}_{\mathrm{s2}})$ depends on how many *k*-mers fall below *σ*. It has been empirically observed (Kuang, et al., 2005) that with a typical choice of *σ*, one *k*-mer in the original sequence translates into m=1, 2 slightly different *k*-mers at the same position. It can then be shown that the worst case complexity of computing a kernel value for a pair of proteins of lengths l_1 and l_2 is $\mathrm{O}\left(\mathrm{k}^{\mathrm{m}+1}|\Sigma|^{\mathrm{m}}(\mathrm{l}_1+\mathrm{l}_2)\right)$. In practice, however, we usually achieve much lower complexity. We again refer to Kuang, et al., 2005 for a more detailed analysis.

SOM Section 3: Multi-class classification schemes

One-Against-All (Allwein and Singer, 2000). Given a set of *n* classes, *n* different binary classifiers are employed such that each classifier discriminates between the positive training instances belonging to one class and the negative training instances belonging to the remaining *n-1* classes. The classification result is the output of the classifier that generates the highest value.

Ensemble of Nested Dichotomies (ENDs) (Frank and Kramer, 2004). A set of twenty randomly composed nested dichotomies (NDs) (Fox, 1997), represented as binary trees. Each internal node of the tree stores one binary classifier and a set of corresponding classes. The root node contains the entire set of classes and learns to separate it into two subsets – a positive and a negative subset. The two successor nodes of the root inherit two subsets and the procedure is repeated until the leaf node is reached. The number of leaf nodes corresponds to the number of localization classes. The result of an END is the average over the estimates obtained from the individual trees.

Ensemble of Class Balanced Nested Dichotomies (ECBNDs) (Dong, et al., 2005). While ENDs sample from a space of all possible tree structures, ECBNDs sample from a space of class-balanced tree structures and built an ensemble of balanced trees. Each internal node in a class-balanced binary tree has two equal-sized subsets to pass to both its successor nodes, which limits the number of possible sets of classes a node can inherit. As a result, the number of possible CBDNs is always smaller than the number of possible NDs.

Ensemble of Data Balanced Nested Dichotomies (EDBNDs) (Dong, et al., 2005). DBNDs are built by randomly assigning classes to two subsets until the number of instances in one of the subsets exceeds half the total amount of instances in the parental node. The two data-balanced subsets are then passed to the successor node. Thus, the heavily populated classes are located high up in the tree structure making the ensemble of possible DBNDs (EDBNDs) biased towards populous classes. However, it has been shown in (Dong, et al., 2005) that the accuracy of EDBNDs is comparable to that of ENDs and ECBNDs on the UCI dataset (Blake and Merz, 1998). This was the reason for investigating this approach on our data.

Nested Dichotomies of a Fixed structure. The knowledge of general pathways of protein sorting was used to design hierarchical trees of a fixed architecture for archaeal, bacterial and eukaryotic proteins (Fig. 1). The difference to ENDs is essentially that we use biological knowledge to define a single ND, instead of randomly creating multiple random NDs and then averaging.

SOM Section 4: Size increase of the training sets.

It has been shown that larger data sets improve SVM performance through increased coverage of the sequence space (Webb and Yu, 2004). Moreover, SVM performance can also be improved through training on sequence redundant sets. Therefore, we allowed a certain degree of homology within each of the training sets of archaeal, bacterial and eukaryotic proteins and thus increased the size of our training data considerably, by almost a factor of 4.

Outline of the algorithm: (1) Start with the homology reduced set and align it against all proteins extracted from SWISS-PROT (Bairoch and Apweiler, 2000) by a pairwise BLAST (Altschul, et al., 1997) (e.g. BLAST2 at E-value<10^4 in our case); (2) Compile HSSP-values (Rost, 1999; Sander and Schneider, 1991) for each pair of aligned sequences. (3) Find all structural homologs to the sequences in the homology reduced set at HSSP-value≤60; (4) Align all sequences found in the previous step against each other by a pairwise BLAST; (5) Find all pairs that are structural homologs at HSSP-value≤60; (6) Remove sequences from the previous step that have HSSP-value>0 to more than one sequence in the homology reduced set.

SOM Section 5: Analysis of k-mers important for endoplasmatic reticulum association.

Although the profile kernel computes dot products implicitly via a kernel trick, the normal vector of the separating hyperplane can be made explicit (Leslie, et al., 2004). This normal vector defines a weight for each k-mer, indicating its contribution to the final classification of a protein: the higher the absolute value, the more the k-mer contributes to the classification of a protein.

However, the biological implications of such a vector are limited (examples given below) and many technical issues would have to be solved before approaching a systematic analysis of our models with respect to new localization signals. (The final class of a protein is the result of many SVMs, each with different normal vectors, to give only one example.)

Nevertheless, for a proof of principle, we computed the explicit normal vector of the SVM separating soluble proteins of the endoplasmatic reticulum (ER; 26 proteins) from those that are secreted or reside in the Golgi apparatus (non-ER; 2318 proteins; Fig. 1). Necessary programs were available in the profile kernel package. Each of the 64 M dimensions of this vector determined the weight of one particular k-mer ($\text{<alphabet size>}^{\text{<k-mer length>}}= 20^6$=64 M; weights ranged from 0.6 to -1.4). We manually analyzed the 100 k-mers with the highest positive weight, i.e. having the highest impact on the classification as ER-associated (weight range: 0.6 – 0.4). A majority came from the C-terminal domain of Calreticulin proteins (e.g. CALR_BOVIN) and only consisted of aspartic and glutamic acids (e.g. EDEDDE, DDEDDE, DEEDEE, …), rendering the domain very acidic. Their high weights can be explained by frequent occurrence in the 26 ER training proteins and absence in the support vectors of non-ER proteins. Indeed, due to our including slightly homologous proteins in the training set, the Calreticulin family was a little overrepresented among ER proteins (6 proteins). The fact that those k-mers weighed higher than other parts of Calreticulin proteins, however, indicated their particular importance for ER localization. Several experimental studies confirmed this hypothesis (Villamil Giraldo, et al., 2010).

A second striking class of high scoring k-me rs consisted of leucin stretches (e.g. LLLLLL, LLLLLA). They could be found in the N-terminal regions of three diverse proteins, namely LDLR chaperone MESD (MESD_HUMAN), GDP-fucose protein O-fucosyltransferase 1 (OFUT1_RAT) and Orexin (OREX_RAT) and are all part of known or putative signal peptides (e.g. [Sakurai, et al., 1999]). However, to our knowledge, exactly this leucin repeat has so far not been identified as a crucial component of the signal and might therefore be a good target for further experimental analyses.

Curiously, the most well-known ER signaling motif, the KDEL retention sequence, was not among the top scoring k-mers despite being largely present in our dataset: 10 of 26 ER proteins ended with the sequence KDEL or HDEL. Further analyses, however, revealed the signal in a different way: There were $2*20^3$ = 800 possible k-mers ending with HDEL or KDEL. 760 of them had a positive weight, only 36 were slightly negative (4 had weight 0.0). This illustrates the need for better methods to detect conserved signals and the limits of the current profile kernel: the location of the motif is important for its function; however it can only partially be captured by our feature space. Many k-mers ending with HDEL or KDEL were also present in non-ER proteins, but not necessarily at the C-terminus.

Table SOM_1: Number of proteins in sequence unique data sets used for the development of LocTree2

Localization	*Eukaryota*	*Bacteria*	*Archaea*
Chloroplast	133	-	-
Chloroplast membrane	11	-	-
Cytosol	220	179	41
Endoplasmic reticulum	10	-	-
Endoplasmic reticulum membrane	65	-	-
Extra-cellular space	596	82	5
Fimbrium	-	16	-
Golgi apparatus	3	-	-
Golgi apparatus membrane	17	-	-
Mitochondria	140	-	-
Mitochondria membrane	87	-	-
Nucleus	320	-	-
Nucleus membrane	5	-	-
Outer membrane	-	6	-
Plasma membrane	40	144	13
Periplasm	-	52	-
Peroxisome	6	-	-
Peroxisome membrane	2	-	-
Plastid	14	-	-
Vacuole	3	-	-
Vacuole membrane	10	-	-
	1682	479	59

The table displays the number of sequences per localization in the sequence unique sets of eukaryotic, bacterial and archaeal proteins. We only used experimentally determined subcellular localization annotations from SWISS-PROT release 2011_04 (Methods).

Table SOM_2: Number of proteins in sequence unique independent test sets

Localization	*New SWISS-PROT*		*LocDB*	
	Bacteria	*Eukaryota*	*A. thaliana*	*H. sapiens*
Chloroplast	-	-	-	-
Cytosol	10	5	2	58
Endoplasmic reticulum (ER)	-	-	3	8
ER membrane	-	3	-	1
Extra-cellular space	8	15	8	14
Fimbrium	1	-	-	-
Golgi apparatus	-	-	1	6
Golgi apparatus membrane	-	1	-	-
Mitochondria	-	3	4	37
Mitochondria membrane	-	3	-	-
Nucleus		15	3	29
Periplasm	3	-	-	-
Plasma membrane	6	5	10	43
Peroxisome	-	-	3	-
Vacuole	-	-	9	5
Vacuole membrane	-	2	-	-
	28	52	43	201

In this table we show the number of sequences per localization in the sequence unique sets of bacterial and eukaryotic SWISS-PROT proteins added between releases 2011_04 and 2012_02 and of *Arabidopsis thaliana* and *Homo sapiens* proteins derived from LocDB (Rastogi and Rost, 2011; Methods). The data sets contained no sequence pairs with HSSP-value>0 and no protein sequences with HSSP-value>5 to any of the sequences used for the development of our prediction method. Note: LocDB annotations of subcellular localization of *A. thaliana* proteins do not discriminate between non-membrane/membrane compartments (with the exception of plasma membrane).

Table SOM_3: Number of proteins in sequence unique sets used for the additional comparison with Loctree

Localization	*Bacteria*	*Eukaryota*
Chloroplast	-	11
Cytosol	22	22
Extra-cellular space	20	84
Mitochondria	-	24
Nucleus	-	22
Organelles	-	21
Periplasm	3	-
	45	184

The table displays bacterial and eukaryotic data sets of protein sequences with localization annotations added to Swiss-Prot after 2005. These sets were used for the performance comparison of LocTree2 to LocTree. The sets sequence redundancy reduced internally (HVAL<0 and BLAST2 EVAL≤10-3 over alignments of >=35 residues length; Methods) and to the training sets of LocTree.

Table SOM_4: Evaluation of prediction accuracy and time of various String Kernels

Method *Performance*		*String Subsequence Kernel* (Lodhi, et al., 2002)	*Mismatch Kernel* (Leslie, et al., 2004)	*Profile Kernel* (Kuang, et al., 2004; SOM Section 2)
Archaea	*Q(3)*	98 ± 3	98 ± 3	98 ± 3
Bacteria	*Q(6)*	89 ± 2	89 ± 2	94 ± 2
Eukaryota	*Q(18)*	70 ± 1	Not available*	83 ± 1

Data set: 10-fold cross-validated training sets of 59 archaeal, 479 bacterial and 1682 eukaryotic proteins (Table SOM_1, cross-validation described in Methods and SOM Section 1).

Performance measures: Q(n), overall prediction accuracy for a given hierarchy (Methods, *Qn* is a 3-state value for archaea, 6-state value for bacteria and 18-state value for eukaryota); Note: the mismatch kernel was not able to produce results on our largest data sets of eukaryotic proteins within a reasonable amount of time.

The averages over all training sets are reported. The multi-class classification approach used was One-against-all (Allwein, et al., 2000; SOM Section 3).

Table SOM_5: Evaluation of prediction accuracy and time of various multi-class classification techniques

Method *Performance*		*One-against-all* (Allwein, et al., 2000)	*ENDs* (Kramer and Frank, 2004)	*ECBNDs (Dong, et al., 2005)*	*EDBND (Dong, et al., 2005)*	*Fixed structure* (Main paper; Fig. 1)
Archaea	*Q(3)*	98 ± 3	98 ± 3	98 ± 3	98 ± 3	98 ± 3
	Speed	$60 \cdot 10^3$	$6.8 \cdot 10^3$	$8.9 \cdot 10^3$	$8.4 \cdot 10^3$	$40 \cdot 10^3$
Bacteria	*Q(6)*	94 ± 2	97 ± 2	97 ± 2	97 ± 2	97 ± 2
	Speed	$16.5 \cdot 10^3$	$2 \cdot 10^3$	$2.6 \cdot 10^3$	$2.7 \cdot 10^3$	$15 \cdot 10^3$
Eukaryota	*Q(18)*	83 ± 1	88 ± 1	88 ± 1	88 ± 1	88 ± 1
	Speed	$4 \cdot 10^3$	$0.2 \cdot 10^3$	$0.3 \cdot 10^3$	$0.3 \cdot 10^3$	$6.1 \cdot 10^3$

Data set: 10-fold cross-validated training sets of 59 archaeal, 479 bacterial and 1682 eukaryotic proteins (Table SOM_1, cross-validation described in Methods and SOM Section 1).

Methods: please refer to SOM Section 3 for methods description.

Performance measures: ***Q(n)*** used as in Table SOM_3; ***Speed***, the number of protein sequences processed per minute on a Dell M605 machine with a Six-Core AMD Opteron processor (2.4 GHz, 6MB and 75W ACP) running on Linux.

The averages over all training sets are reported. The kernel function used was the Profile Kernel (SOM Section_2).

Table SOM_6: LocTree2 on non-redundant test sets of 479 bacterial 1682 and eukaryotic proteins

Localization	*Nprot*	*Acc*	*Cov*	*gAv*
Cytosol	179	87 ± 6	92 ± 5	89 ± 6
Extra-cellular	82	63 ± 12	73 ± 11	68 ± 11
Fimbrium	16	83 ± 25**	31 ± 32**	51 ± 31
Periplasm	52	75 ± 16	58 ± 16	66 ± 16
Plasma membrane	144	96 ± 4	95 ± 4	95 ± 5
Q(6) – Bacteria		84 ± 4		
Chloroplast	133	44 ± 13	29 ± 9	36 ± 7
Chloroplast membrane	11	38 ± 48**	27 ± 30	32 ± 27
Cytosol	220	45 ± 8	44 ± 8	44 ± 6
ER membrane	65	44 ± 15	42 ± 14	43 ± 11
Extra-cellular	596	80 ± 4	91 ± 3	85 ± 4
Golgi membrane	17	42 ± 33	29 ± 24	35 ± 19
Mitochondria	140	45 ± 10	46 ± 10	45 ± 7
Mitochondria membrane	87	60 ± 15	44 ± 13	51 ± 10
Nucleus	320	67 ± 6	77 ± 6	72 ± 6
Plasma membrane	40	68 ± 22	48 ± 19	57 ± 18
Plastid	14	50 ± 50	21 ± 28	33 ± 21
Q(18) – Eukaryota		65 ± 3		

Abbreviations used: *Nprot*, the number of proteins with known localization; *Acc*, accuracy; *Cov*, coverage; *gAv*, geometric coverage of *Acc* and *Cov*; *Q(n)*, overall prediction accuracy. Standard errors were estimated by bootstrapping (Methods). Note 1: *Q(n)* is a six-state value for bacteria, i.e. the overall accuracy for classification in one of six localization classes, and an eighteen-state value for eukaryota. Note 2: Only performances for localization classes containing more than ten proteins are reported.

** = unrealistic upper or lower bound given by the standard error due to the small data set size.

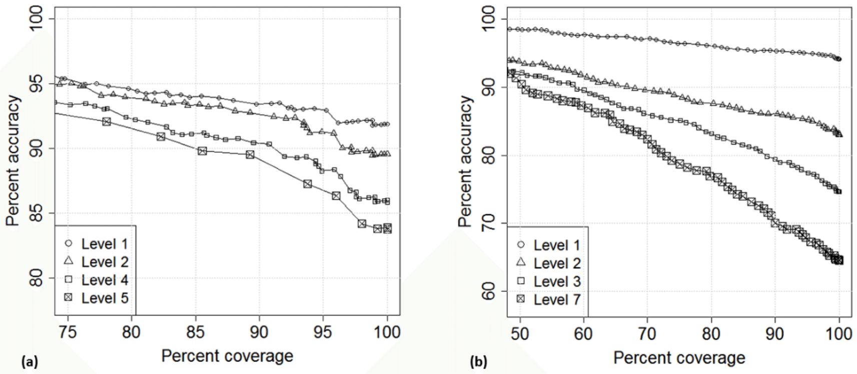
Table SOM_7: LocTree2 on non-redundant test set of 59 archaeal proteins

Localization	*Nprot*	*Acc*	*Cov*	*gAv*
Cytosol	41	100*	100*	100*
Extra-cellular	5	100*	100*	100*
Plasma membrane	13	100*	100*	100*
Q(3)		100*		

Abbreviations used as for Table SOM_6. Note: *Q(n)* is a three-state value, i.e. the overall accuracy for classification in one of three localization classes.

* = overoptimistic estimate due to the small data set size.

Figure SOM_1: LocTree2 performance on cross-validated test sets of 479 bacterial and 1682 eukaryotic proteins.



The curves show the overall prediction accuracy on cross-validated sets of sequence unique bacterial and eukaryotic proteins used for the development of LocTree2 (Table SOM_1, Methods). (a) Bacteria: The level 1 accuracy represents the overall two-state accuracy of classifying proteins into cytoplasmic and non-cytoplasmic classes (Fig. 1b). For example, at 90% coverage, the prediction accuracy was around 94%. The overall accuracy declined at lower levels in the hierarchical tree. At Level 2 node that separates proteins into cytosolic, plasma membrane and non-plasma membrane classes, the overall accuracy decreased to 93% at 90% coverage. For the purpose of simplification, the curve for Level 3 predictions is not here provided. Level 4 accuracy includes the accuracies of the cytosolic, plasma membrane, periplasmic space, outer membrane and non-outer membrane classes; it was 91% at 90% coverage. The difference in accuracy between Level 1 and Level 5 predictions that separate proteins in one of six subcellular localization classes was 9%. (b) Eukaryota: the decision node at Level 1 (Fig. 1c) separated membrane spanning proteins from those not associated with membranes. At 80% coverage the accuracy was 96%. Similarly to Fig. 1a, the prediction accuracy declined with the depth of the classification tree. The predictions at Level 2 nodes, where non-membrane and transmembrane proteins are separated into secretory pathway/non-secretory pathway proteins, were made at a lower level of 88% accuracy at 80% coverage. Level 3 nodes separated proteins into eight classes at 83% accuracy and 80% coverage. The prediction into one of eighteen localization classes at Level 7 nodes was performed at a significantly lower accuracy of around 77% at 80% coverage. The performances at Levels 4-6 are explicitly not provided in order to simplify.

Table SOM_8: Performance comparison on LocDB data with several evidences

	LocDB (>2 publications)					
	A. thaliana = 10 proteins			*H. sapiens = 18 proteins*		
Method	*Q(9)*	*Q(8)*	*Q(6)*	*Q(8)*	*Q(7)*	*Q(6)*
LocTree2	**40±34**	**44±37**	**60±48**	**67±27**	**67±27**	**85±24**
CELLO v. 2.5	**40±34**	-	-	61±28	-	-
WoLF PSORT	**40±34**	-	-	61±27	-	-
MultiLoc2	-	22±41	-	-	56±26	-
LOCtree	-	-	**60±48**	-	-	**85±24**

Data set: 10 *Arabidopsis thaliana* and 18 *Homo sapiens* sequence-unique proteins from the LocDB database with localization annotations supported by more than two publications. Both data sets were redundancy reduced (HVAL<0; Methods) with respect to each other and to SWISS-PROT 2011_04.

Performance measure: in each column, the highest achieved overall accuracy *Q(n)* is marked in bold letters; values ± standard error (Methods)

Table SOM_9: Estimating the influence of sequencing mistakes

Method	*Full length*	*30N removed*	*30C removed*	*1/3 randomly removed*
LocTree2	**62 ± 2**	**54 ± 2**	**60 ± 2**	**53 ± 2**
CELLO v. 2.5	56 ± 2	35 ± 2	47 ± 2	48 ± 2
WoLF PSORT	56 ± 2	40 ± 2	52 ± 2	49 ± 2

Data set: combined set of all sequence-unique eukaryotic protein sequences extracted from SWISS-PROT and LocDB databases (Tables SOM_1 and SOM_2).

Methods: We estimated and compared the effect of sequencing errors on the performance of LocTree2 and its competitors. Three data sets were used: *30N removed*, the first thirty N-terminal amino acids were cleaved off for all proteins; *30C removed*, the first thirty C-terminal amino acids were cleaved off for all proteins; *1/3 randomly removed*, amino acid positions were randomly picked and cleaved off in-silico until two thirds of the protein sequence remained, which was used to predict localization. *Full length*, is the performance on full length protein sequences and is shown here for comparison.

Performance measure: in each column, the highest achieved overall accuracy is marked in bold letters.

References for Supporting Online Material

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, *Journal of molecular biology*, 403-410 (1990)

S.F. Altschul, *et al, Nucleic acids research,* 25, 3389-3402. (1997)

E.L. Allwein, and Y. Singer, *J. Machine Learning Research*, vol. 1, 113-141 (2000)

A. Bairoch and R. Apweiler, R. *Nucleic acids research*, **28**, 45-48. (2000)

Berman, H.M.*, et al.* (2000) The Protein Data Bank, *Nucleic acids research*, **28**, 235-242.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H.Weissig, I.N. Shindyalov, P.E. Bourne, *Nucleic acids research*, **28**, 235-242 (2000)

C. Blake, C.Merz, *University of California, Irvine, Dept. of Inf. and Computer Science* (1998) [www.ics.uci.edu/~mlearn/MLRepository.html]

C. Cortes, V. Vapnik, *Machine Learning*, 273-297 (1995)

L. Dong, E. Frank and S. Kramer*, PKDD ,* 84-95 (2005)

J. Fox, *Applied Regression Analysis, Linear Models, and Related Methods.* Sage Publication. (1997)

G. Holmes, A. Donkin, I.H. Witten, *Proc Second Australia and New Zealand Conference on Intelligent Information Systems*, 357-361 (1994)

S. Kramer, E. Frank, *Proceedings of ICML* (2004)

R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, C. Leslie, *Proc IEEE Comput Syst Bioinform Conf*, 152-160 (2004)

H, Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins, *The Journal of Machine Learning Research, vol. 2, 419-444,* (2002)

C. Leslie, E. Eskin, A. Cohen, J. Weston W.S. Noble, *Bioinformatics*, 20(4): 467-476, 2004.

J.C. Platt, *Advances in Kernel Methods - Support Vector Learning* (1998)

S. Rastogi, B. Rost, *Nucleic acids research*, D230-234 (2011)

B. Rost, *Protein engineering*, 85-94 (1999)

T. Sakurai, T. Moriguchi, K. Furuya, N. Kajiwara, T. Nakamura, M. Yanagisawa, K. Goto, *The Journal of biological chemistry,* **274**(25):17771-17776 (1999)

C. Sander, R. Schneider, *Proteins*, 56-68 (1991)

A.M. Villamil Giraldo, M. Lopez Medus, M. Gonzales Lebrero, R.S. Pagano, C.A. Labriola, L. Landolfo, J.M. Delfino, A.J. Parodi, J.J. Caramelo, The Journal of biological chemistry, 285(7):4544-4553 (2010)

G. Webb, X.E. Yu, Advanced in Artificial Intelligence, Proceedings of 17th Australian Joint Conference on AI, Lecture Notes in Artificial Intelligence (2004)

2.3 References

1. Andrade MA, O'Donoghue SI, Rost B: **Adaptation of protein surfaces to subcellular location.** *Journal of molecular biology* 1998, **276:**517-525.
2. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, et al: **Prediction of human protein function from post-translational modifications and localization features.** *Journal of molecular biology* 2002, **319:**1257-1265.
3. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y: **Automatic prediction of protein function.** *Cellular and molecular life sciences : CMLS* 2003, **60:**2637-2650.
4. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.
5. Foster LJ, de Hoog CL, Zhang Y, Xie X, Mootha VK, Mann M: **A mammalian organelle map by protein correlation profiling.** *Cell* 2006, **125:**187-199.
6. Li S, Ehrhardt DW, Rhee SY: **Systematic analysis of Arabidopsis organelles and a protein localization database for facilitating fluorescent tagging of full-length Arabidopsis proteins.** *Plant physiology* 2006, **141:**527-539.
7. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26:**1608-1615.
8. Briesemeister S, Blum T, Brady S, Lam Y, Kohlbacher O, Shatkay H: **SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins.** *Journal of proteome research* 2009, **8:**5363-5366.
9. Chou KC: **Prediction of protein cellular attributes using pseudo-amino acid composition.** *Proteins* 2001, **43:**246-255.
10. Hoglund A, Donnes P, Blum T, Adolph HW, Kohlbacher O: **MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition.** *Bioinformatics* 2006, **22:**1158-1165.
11. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor.** *Nucleic acids research* 2007, **35:**W585-587.
12. Nair R, Rost B: **Inferring sub-cellular localization through automated lexical analysis.** *Bioinformatics* 2002, **18 Suppl 1:**S78-86.
13. Nair R, Rost B: **Mimicking cellular sorting improves prediction of subcellular localization.** *Journal of molecular biology* 2005, **348:**85-100.
14. Park KJ, Kanehisa M: **Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs.** *Bioinformatics* 2003, **19:**1656-1663.
15. Reinhardt A, Hubbard T: **Using neural networks for prediction of the subcellular location of proteins.** *Nucleic acids research* 1998, **26:**2230-2236.
16. Wrzeszczynski KO, Rost B: **Annotating proteins from endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes.** *Cellular and molecular life sciences : CMLS* 2004, **61:**1341-1353.
17. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization.** *Proteins* 2006, **64:**643-651.
18. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S: **Feature-based prediction of non-classical and leaderless protein secretion.** *Protein engineering, design & selection : PEDS* 2004, **17:**349-356.

19. Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A: **Prediction of membrane-protein topology from first principles.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105:**7177-7181.

20. Kall L, Krogh A, Sonnhammer EL: **An HMM posterior decoder for sequence feature prediction that includes homology information.** *Bioinformatics* 2005, **21 Suppl 1:**i251-257.

3 LocTree3: improved prediction of protein cellular sorting

3.1 Preface

LocTree2 [1], a method described in Chapter 2, accurately predicts proteins in the so far largest number of localization classes using machine learning. An independent benchmark study of Mooney *et al*. [2] proved LocTree2 to be a successor and/or compliment of other state-of-the-art methods. Another study, of Imai and Nakai [3], suggested a simple homology-based inference, *i.e.* annotation transfer from experimentally annotated sequence homologs, to perform *on par with* or better than advanced machine learning methods.

In this publication, we compared the performance of LocTree2 (*de novo*-based predictions) with that of PSI-BLAST [4] (homology-based inference) on cross-validated sequence-unique development data of Loctree2. We found that PSI-BLAST could, indeed, significantly outperform LocTree2 for about half of the proteins in our set, for which homologous proteins of known localization were available. For other proteins, the homology-based inference was not possible. Thus, we argued that whole proteome annotations using sequence homology only are rather limited, and suggested a new protocol that combines homology-based inference if available with *de novo* predictions, otherwise. The resulted method, LocTree3, outperformed its predecessor LocTree2 by remarkable 25%. We applied LocTree3 to the proteomes of all entirely sequenced organisms and showed that in human, for instance, localization for 23% of all proteins can only be inferred *de novo* (for yeast this number is 32%, *A. thaliana* 39% and archaea *A. pernix* 92%). Furthermore, this publication initiates a discussion that, in our opinion, is of significant importance in the field, as it addresses questions such as the reliability of experimental data for localization in current databases and of interpretation of the computational prediction results.

The study design was conceived by me, Henrik Nielsen and Burkhard Rost. I carried out necessary background search. The initial evaluation of the performances of homology-based and *de novo* predictions was performed by students of the “Protein Prediction II” practical course (winter term 2013/14) under my and Maximilian Hecht’s guidance. The combination of two sources of prediction into LocTree3 and the method’s subsequent evaluation was done by me and Burkhard Rost. I programmed LocTree3, while the implementation of the faster version of the Profile Kernel [5, 6] (required for LocTree2) came from Tobias Hamp. LocTree3’s web server was implemented by me, Maximilian Hecht, Timothy Karl and Guy Yachdav. The manuscript was drafted by me and Burkhard Rost.

3.2 Journal article. Goldberg T., Hecht M., Rost B., *et al*. *NAR* 2014; 42:W350-5

W350–W355 Nucleic Acids Research, 2014, Vol. 42, Web Server issue *Published online 21 May 2014*
doi: 10.1093/nar/gku396

LocTree3 prediction of localization

Tatyana Goldberg[1,2,*,†], Maximilian Hecht[1,†], Tobias Hamp[1], Timothy Karl[1], Guy Yachdav[1,3], Nadeem Ahmed[1], Uwe Altermann[1], Philipp Angerer[1], Sonja Ansorge[1], Kinga Balasz[1], Michael Bernhofer[1], Alexander Betz[1], Laura Cizmadija[1], Kieu Trinh Do[1], Julia Gerke[1], Robert Greil[1], Vadim Joerdens[1], Maximilian Hastreiter[1], Katharina Hembach[1], Max Herzog[1], Maria Kalemanov[1], Michael Kluge[1], Alice Meier[1], Hassan Nasir[1], Ulrich Neumaier[1], Verena Prade[1], Jonas Reeb[1], Aleksandr Sorokoumov[1], Ilira Troshani[1], Susann Vorberg[1], Sonja Waldraff[1], Jonas Zierer[1], Henrik Nielsen[4] and Burkhard Rost[1,3,5,6,7]

[1]Department of Informatics, Bioinformatics-I12, TUM, 85748 Garching, Germany, [2]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), 85748 Garching, Germany, [3]Biosof LLC, New York, NY 10001, USA, [4]Center for Biological Sequence Analysis, Department of Systems Biology, DTU, 2800 Lyngby, Denmark, [5]Institute for Advanced Study (TUM-IAS), 85748 Garching, Germany, [6]New York Consortium on Membrane Protein Structure (NYCOMPS) & Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA and [7]Institute for Food and Plant Sciences WZW – Weihenstephan, 85350 Freising, Germany

Received February 13, 2014; Revised April 11, 2014; Accepted April 24, 2014

ABSTRACT

The prediction of protein sub-cellular localization is an important step toward elucidating protein function. For each query protein sequence, LocTree2 applies machine learning (profile kernel SVM) to predict the native sub-cellular localization in 18 classes for eukaryotes, in six for bacteria and in three for archaea. The method outputs a score that reflects the reliability of each prediction. LocTree2 has performed *on par with* or better than any other state-of-the-art method. Here, we report the availability of LocTree3 as a public web server. The server includes the machine learning-based LocTree2 and improves over it through the addition of homology-based inference. Assessed on sequence-unique data, LocTree3 reached an 18-state accuracy *Q*18 = 80 ± 3% for eukaryotes and a six-state accuracy *Q*6 = 89 ± 4% for bacteria. The server accepts submissions ranging from single protein sequences to entire proteomes. Response time of the unloaded server is about 90 s for a 300-residue eukaryotic protein and a few hours for an entire eukaryotic proteome not considering the generation of the alignments. For over 1000 entirely sequenced organisms, the predictions are directly available as downloads. The web server is available at http://www.rostlab.org/services/loctree3.

INTRODUCTION

Many experimental methods annotate protein localization, enriching resources such as SWISS-PROT (1). However, even for the well-studied yeast, the experimental data are not nearly complete (2,3). Bridging the sequence-annotation gap (4) for localization, therefore, calls for cheaper and faster *in silico* approaches (5,6). Many machine learning methods predict the native localization of a protein from its amino acid sequence; among the best known are CELLO (7), WoLF PSORT (8), YLoc (9) and PSORTb (10). A recent study suggested homology-based inference to outperform machine learning (11). Homology-based inference proceeds as follows: build a data set with all proteins of known localization, run a simple pairwise BLAST (12) against this set, and predict the localization of the first hit.

LocTree2 predicts a single localization for all proteins in all domains of life through machine learning (13). The method implements a hierarchical system of Support Vector Machines (SVMs) to imitate the cascading mechanism of cellular sorting (14). An independent, recent benchmark proved LocTree2 to be an excellent successor and/or complement to other top-of-the-line prediction methods (15) in situations in which no experimental information is available for the query protein or its homologs.

Here, we introduce LocTree3. It provides the web server front end for LocTree2, and improves over LocTree2 by including information about homologs if available. Thereby, LocTree3 combines 'the best of both worlds', employing homology when possible and machine learning otherwise. The

[*]To whom correspondence should be addressed. Tel: +49 89 2891 7850; Fax: +49 89 2891 9414; Email: goldberg@rostlab.org
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

major steps of improvement are as follows: (i) inclusion of annotation transfer from close homologs with experimentally annotated localization through PSI-BLAST (12); (ii) runtime reduction of LocTree2 by using a new fast implementation of the SVM profile kernel (16,17); (iii) Gene Ontology (18) annotations for prediction results; (iv) caching of the results for faster processing of the repeated searches (19,20).

MATERIALS AND METHODS

Data

The number of proteins with experimental annotation for a single localization in SWISS-PROT release 2011_04 was 34 583 for eukaryotes (18 localization classes, visualized in Figure 2), 4765 for bacteria (six classes: cytosol, plasma membrane, periplasmic space, outer membrane, fimbrium and extra-cellular) and 237 for archaea (three classes: cytosol, plasma membrane and extra-cellular). LocTree2 was developed on sequence-unique subsets with 1682 eukaryotic, 479 bacterial and 79 archaeal proteins (Supplementary Table S1, Supporting Online Material). Sequence-redundancy was reduced at HVAL ≤ 0 (21,22) through UniqueProt (23). This is commonly done because the bias in data sets from sequence similarity often overestimate performance (24). However, in order to assess the power of homology-based inference, we had to accept some redundancy because homology-based inference performed below the level of random across sequence-unique proteins (Supplementary Table S2). We accomplished this by running the sequence-unique 1682 eukaryotic proteins against all experimentally annotated proteins, i.e. against the same release of SWISS-PROT putting the redundancy back in to enable PSI-BLAST lookups. For 995 of the 1682, PSI-BLAST found a non-trivial (removal of query protein) at E-value $\leq 10^{-3}$ (25,26); for 687 it did not.

For further testing, we added three new data sets. We collected all proteins for which experimental annotations had been added between releases 2011_04 and 2013_11. We redundancy reduced those at HVAL ≤ 0. This gave the sets New2013_hval0 (273 for eukaryotes, 57 for bacteria). Additional redundancy reduction to LocTree3 development data provided too small sets (32 eukaryotic and two bacterial proteins) for reliable performance estimates. Next, we simulated the question 'how well the method will perform on the next 1000 new proteins?' by simply monitoring all proteins added since we began collecting the data for this manuscript, i.e. the proteins added since 2013_11 (New2014 with 198 eukaryotic proteins and too few in bacteria to proceed). Finally, we investigated a third set with all human proteins (Supplementary Table S3). We deliberately kept the 'redundancy' in this set that exists on the level of an organism. Note that throughout we have considered only proteins with single experimental annotations. Our preliminary analysis of proteins with multiple annotations suggested these to constitute a small set of proteins with many problematic annotations (Supplementary Section S1).

Methods

(1) **Homology-based inference:** We transferred localization annotations by homology through PSI-BLAST (12). For all proteins with experimentally known localization, we generated PSI-BLAST profiles using an 80% non-redundant database combining UniProt (1) and PDB (27) with two iterations and *E*-value $\leq 10^{-3}$. These profiles were then aligned against all proteins with experimental annotation of a single localization in SWISS-PROT release 2011_04. PSI-BLAST hits to the input protein were excluded.

(2) **LocTree2** (13) utilizes a hierarchical system of SVMs. At all levels of the tree are binary decisions, which are made by searching through proteins of annotated localization with short stretches of *k*-consecutive residues ($k = 3$ for archaea, 5 for bacteria and 6 for eukaryota). The most informative *k*-mer hit decides on 'left or right' for each fork in the tree until reaching a leaf, i.e. the final predicted localization class.

(3) **LocTree3:** Our final method, LocTree3, combines PSI-BLAST and LocTree2 in the settings where they perform best. A single parameter chooses: homology-based inference, if a profile-2-sequence PSI-BLAST hits at *E*-value $\leq 10^{-3}$, else: LocTree2 ('Results' section and Supplementary Figures S1 and S2).

(4) **Public methods (CELLO 2.5, WoLF PSORT, YLoc, PSORTb 3.0):** We compared LocTree3 to four publicly available leading prediction methods: CELLO 2.5 (7), WoLF PSORT (8), YLoc (9) and PSORTb 3.0 (10). If WoLF PSORT or CELLO 2.5 predicted multiple locations, and one of those was correct, we always considered the prediction fully correct. Furthermore, these two methods distinguish cytoskeleton and cytoplasm; here, we considered both as cytosolic. Because no method other than LocTree2/3 distinguishes between membranes other than the cell membrane in eukaryotes, we merged these two classes, i.e. treated nuclear and nuclear-membrane proteins as identical. Plastid and chloroplast proteins were also merged into one class for a comparison of LocTree3 to other methods. For a comparison with CELLO 2.5 and PSORTb 3.0 we combined bacterial secreted and fimbrium proteins into one class and differentiated between Gram-positive and Gram-negative proteins according to Yu *et al.* (10).

Reliability index

The reliability of a prediction is given through a reliability index ranging from 0 (weak prediction) to 100 (confident prediction). For LocTree2, the reliability indices are taken directly from its output. For homology-based inferences from PSI-BLAST, the reliability index was compiled as a simple function of the percentage pairwise sequence identity (PIDE) with a threshold at the saturation of PIDE ≤ 20 (Supplementary Figure S1).

Performance evaluation

The performance for a single localization class L was expressed using accuracy (often also referred to as precision)

Table 1. Performance for LocTree3 and its sources

Method	Eukaryota $Q18$ (Equation (3))			Bacteria $Q6$ (Equation (3))		
	Set2011_hval0 (1682)[*4]	Without PSI-BLAST hits (687) [*4]	With PSI-BLAST hits (995)[*4]	Set2011_hval0 (479)[*5]	Without PSI-BLAST hits (277)[*5]	With PSI-BLAST hits (202)[*5]
PSI-BLAST[*1]	55 ± 3	na	**93 ± 2**	40 ± 5	na	**94 ± 4**
LocTree2[*2]	65 ± 3	**61 ± 5**	67 ± 4	84 ± 4	**84 ± 5**	83 ± 6
LocTree3[*3]	**80 ± 3**			**89 ± 4**		

Note: '±' values refer to standard errors (Equation (4)); bold face: 'winner in each column'.
[*1] PSI-BLAST: simple look-up of localization from proteins with known localization, excluding self-hits.
[*2] LocTree2: *de novo* machine learning-based prediction (cross-validated).
[*3] LocTree3: takes PSI-BLAST if available and LocTree2, otherwise.
[*4]Eukaryotic 'Set2011_hval0': 1682 sequence-unique eukaryotic proteins with experimental localization annotation from SWISS-PROT release 2011_04; for 995 of those, PSI-BLAST found hits at E-value $\leq 10^{-3}$ in the set of all annotations of release 2011_04, for 687 it did not.
[*5]Bacterial 'Set2011_hval0': SWISS-PROT release 2011_04 had localization annotations for 479 sequence-unique bacterial proteins; for 202 PSI-BLAST identified hits in the remainder of annotated proteins in 2011_04, for 227 it did not.

and coverage (often also referred to as recall):

$$\mathrm{Acc(L)} = 100 \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{PF}} \tag{1}$$

$$\mathrm{Cov(L)} = 100 \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

with: TP, the true positives (i.e. the number of proteins predicted and observed in localization L); FP, the false positives (i.e. the number predicted in L and observed in non-L); FN, the false negatives (i.e. the number observed in L and predicted in non-L). We measured the overall performance by the n-state accuracy Qn:

$$Qn = \frac{\text{number proteins correctly predicted in } n \text{ classes}}{\text{total number proteins observed in } n \text{ classes}} \tag{3}$$

Standard errors were estimated over 1000 bootstrap sets, i.e. randomly select 15% of proteins without replacement from the original data set (in our experience this non-standard procedure yields more long-lived estimates). For each bootstrap set, the performance x_i (e.g. accuracy) is estimated through its difference from the overall performance $\langle x \rangle$. These 1000 estimates provided the standard deviation of x_i with the typical standard error, where n is the number of bootstrap sets:

$$\text{Standard deviation}\,(\sigma) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \langle x \rangle)^2}{n}} \tag{4}$$

$$\text{Standard error} = \frac{\sigma}{\sqrt{n-1}}$$

Runtime analysis

For sequences with pre-calculated PSI-BLAST profiles the LocTree2 runtime was measured on a Dell M605 machine with a Six-Core AMD Opteron processor (2.4 GHz, 6MB and 75 W ACP) running on Linux.

RESULTS

LocTree3 balanced PSI-BLAST and LocTree2

Homology-based inference for a protein of unknown localization U implies to find a protein with known localization K that is sequence similar to U (e.g. sim(U,K) > T and U $\neq$ K). We experimented with alternative solutions, but avoided to 'over-optimize'. We simply chose the threshold T to be the standard PSI-BLAST E-value of 10^{-3} (Supplementary Figure S2, Supporting Online Material). This typically gave several hits: choosing the one with highest percentage pairwise sequence identity slightly outperformed taking the hit with best E-value (Supplementary Table S4).

Surprisingly, homology inference outperformed our advanced machine learning tool LocTree2 for half of our original data (995 of 1682 eukaryotic and 202 of 479 bacterial proteins, Table 1). However, when we forced PSI-BLAST to return hits for all proteins, LocTree2 consistently outperformed the PSI-BLAST protocol (Table 1).

These first results suggested a simple protocol: use PSI-BLAST if applicable, LocTree2 if not. We dubbed the method that realized this protocol LocTree3. The combination outperformed both its sources, reaching an overall performance of $Q18 = 80 \pm 3\%$ in classifying eukaryotic proteins in 18 classes (10 non-membrane and 8 membrane classes) and bacterial proteins in six classes at $Q6 = 89 \pm 4\%$ (Table 1). LocTree3 predicted eukaryotic extra-cellular proteins best (Acc: 88% and Cov: 96%), followed by nuclear proteins (Acc: 81% and Cov: 86%; Supplementary Figure S3A, Supplementary Table S5). For bacteria, the prediction of plasma membrane proteins was most accurate (Acc: 96% and Cov: 95%), followed by cytosolic proteins (Acc: 91% and Cov: 90%; Supplementary Figure S3B, Supplementary Table S5).

LocTree3 outperformed other methods

For both eukaryotes and bacteria, LocTree3 significantly outperformed its competitors on all data sets tested (Table 2 and Supplementary Table S6). Finally, we used all experimentally annotated human proteins to benchmark the methods and found LocTree3 again to provide the most accurate predictions (Supplementary Table S7). The complete human set contained 5016 proteins; LocTree3 reached $Q10$

Table 2. Performance comparison for state-of-the-art prediction methods

	Eukaryota *Q*10 (Equation (3))			Bacteria *Q*5 (Equation (3))	
Method	Set2011_hval0 (1682)[*2]	New2013_hval0 (273)[*3]	New2014 (198)[*4]	Set2011_hval0 (479)[*2]	New2013_hval0 (57)[*3]
Cello 2.5[*1]	65 ± 3	64 ± 7	81 ± 7	82 ± 4	70 ± 14
PSORTb 3.0[*1]	-	-	-	57 ± 5	51 ± 15
Wolf Psort[*1]	60 ± 3	65 ± 7	77 ± 7	-	-
YLoc[*1]	60 ± 3	63 ± 7	66 ± 8	-	-
LocTree2	65 ± 3	66 ± 7	**85 ± 6**	86 ± 4	81 ± 11
LocTree3	**81 ± 3**	**73 ± 7**	84 ± 6	**90 ± 3**	**84 ± 11**

Note: '±' values refer to standard errors (Equation (4)); bold face: 'winner in each column'.
[*1] Cello 2.5 (7), PSORTb 3.0 (10), Wolf Psort (8), YLoc (9) as described in 'Materials and Methods' section.
[*2] Set2011_hval0 (as in Table 1): 1682 sequence unique eukaryotic and 479 bacterial proteins used for development of LocTree3.
[*3] New2013_hval0: 273 eukaryotic and 75 bacterial proteins added to SWISS-PROT between releases 2011_04 and 2013_11, sequence homology reduced at HVAL < 0.
[*4] New2014: 198 eukaryotic proteins added to SWISS-PROT between releases 2013_11 and 2014_03 (not redundancy-reduced).

= 89%, followed by YLoc, Cello 2.5 and Wolf Psort with 76, 75 and 71% respectively (Supplementary Table S7). LocTree3 appears best when compared on the same number of classes, and it also is the method that distinguishes in most detail with 18 classes for eukaryotes (compared to 12 for Cello 2.5 and Wolf PSORT; 11 for YLoc).

Reliability index enables users to focus on best predictions

LocTree3 measures the confidence of each prediction through a reliability index (RI) that scales from 0 (low confidence) to 100 (high confidence). Technically, RI reflects the strength of a prediction. Our task as developers was to provide a measure that allows users to translate this strength into estimates for performance. Indeed, our RI strongly correlated with accuracy (Figure 1): when choosing the 50% most strongly predicted eukaryotic proteins, 95% of the predictions were correct (RI > 70, Figure 1: black arrow). For bacterial proteins the same level of accuracy was also reached for about half of all proteins (but at RI > 80, Figure 1: gray arrow). For users not familiar with reliability indices it is important to point out that the choice of the 'top N' does not require knowing the answer. Instead, any user can make this choice for any prediction and can read of Figure 1 what to expect from the choice.

About 90 s runtime without alignment

At this point, the *PredictProtein cache* (19,20) holds >11.7 million pre-computed PSI-BLAST profiles that are quickly retrieved by LocTree3. Due to a recent acceleration of the profile kernel (16,17), the runtime of LocTree2 could be reduced by up to 100 times, such that now an average SVM kernel lookup takes about 90 s for a typical eukaryotic protein (bacteria: 4s, archaea: 2s).

Due to considerable 'start-up' overhead, the runtime increases sub-linearly with the number of queries. This renders the server fit for queries with entire proteomes, typically requiring few minutes for archaeal, <1 h for bacterial and <1 day for eukaryotic proteomes. If the PSI-BLAST profiles have to be created first, runtimes increase manifold, as creating a profile takes 10–500 times longer than running LocTree2. Interested users may download the LocTree3 Debian package from the web server and run it on their machines.

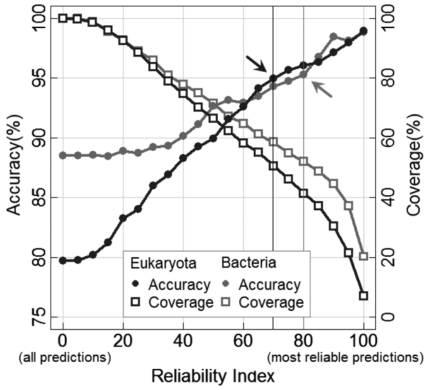
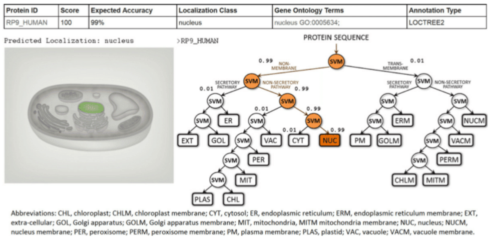


Figure 1. Reliable predictions more accurate. The reliability index (RI) of LocTree3 relates the strength of a prediction to the performance. The curves show the percentage accuracy/coverage ('Materials and Methods' section) for LocTree3 predictions above a given RI. Increasing the RI implies that we look at some subset of all predictions; the subset is given by the curves with squares. For instance, half of all eukaryotic proteins are predicted at RI > 70 (black cross-line). For this top 50%, performance rises from the average *Q*18 = 80% to *Q*18 = 95% (black line with circles, black arrow). Similar values are reached for RI > 80 for bacteria (gray cross-line; note that in this case *Q*6 = 95% is a six-state accuracy as opposed to the 18-state value for eukaryotes).

Prediction workflow

Users submit one or more FASTA-formatted protein sequences. For each sequence, the server first checks for the pre-calculated results in the *PredictProtein cache*. If available, the result is returned immediately (minus queue waiting time); if not, the server retrieves a PSI-BLAST profile through the PredictProtein pipeline (19,20). The profile is used to identify hits in a database of experimentally annotated proteins. If no hits are identified, the profile triggers a *de novo* prediction by LocTree2.



Protein ID	Score	Expected Accuracy	Localization Class	Gene Ontology Terms	Annotation Type
RP9_HUMAN	100	99%	nucleus	nucleus GO:0005634;	LOCTREE2

Predicted Localization: nucleus

Abbreviations: CHL, chloroplast; CHLM, chloroplast membrane; CYT, cytosol; ER, endoplasmic reticulum; ERM, endoplasmic reticulum membrane; EXT, extra-cellular; GOL, Golgi apparatus; GOLM, Golgi apparatus membrane; MIT, mitochondria, MITM mitochondria membrane; NUC, nucleus; NUCM, nucleus membrane; PER, peroxisome; PERM, peroxisome membrane; PM, plasma membrane; PLAS, plastid; VAC, vacuole; VACM, vacuole membrane.

Figure 2. Example output for protein RP9_HUMAN. For every input protein sequence the LocTree3 prediction result contains: (i) protein identifier, (ii) reliability index, (iii) expected accuracy of the prediction, (iv) localization class, (v) GO term(s) and identifier(s) and (vi) source of the prediction. The predicted localization is highlighted in the schematic representation of the cell (here: nucleus). For LocTree2 predictions (shown here), we provide a visualization of the decision tree and the decision path leading to the final prediction. The reliability index is formed through the product of values along the decision path. For PSI-BLAST predictions, we provide a sequence alignment of the query protein to its best hit instead of the tree.

For every query protein, the result contains four basic values: (i) the protein identifier as provided by the user, (ii) the reliability score of a prediction on a 0–100 scale with 100 being the most confident prediction, (iii) single predicted localization class and (iv) GO term(s) and GO identifier(s) matching the predicted class. Every result is supported by the information on whether it comes from a PSI-BLAST homology search or a LocTree2 *de novo* prediction. In case of the former, the web site provides 'per click' on the prediction result the experimental SWISS-PROT annotation of the best hit and its PSI-BLAST alignment to the query protein. In case of the latter, 'the click' on the result will forward to the visual representation of the LocTree2 decision tree and the decision path leading to the final prediction. In addition, every result is supported by a schematic representation of the biological cell highlighting the predicted localization (Figure 2).

Predictions pre-calculated for over 1000 organisms

LocTree3 predictions for over 1000 complete eukaryotic and prokaryotic proteomes are available on the web server (http://rostlab.org/services/loctree3/proteomes/). Predictions are based on sequence sets from the European Bioinformatics Institute (EBI: http://www.ebi.ac.uk/genomes/ and http://www.ebi.ac.uk/reference_proteomes/). The high-throughput annotation and prediction of protein sub-cellular localization allows organism-wide comparisons of protein localization patterns and the reconstruction of evolutionary relations (Goldberg *et al.*, in preparation). Predictions for the newly completed proteomes will be added to the web server on a semi-annual basis.

DISCUSSION

PSI-BLAST has certainly changed the way we do sequence analysis more than any tool (possibly excluding PubMed and Google). Furthermore, this tool has been improving continuously since its first publication in 1997 adding important value beyond that from growing databases (25). LocTree2 uses advanced SVM profile kernels (16). Although it explicitly uses local sequence similarity, LocTree2 arguably falls into the class of *de novo* methods simply because it reaches its predictions through levels of sequence similarity that are not available directly from sequence comparisons. Nevertheless, we found that a simple PSI-BLAST protocol could outperform LocTree2 for about half of the proteins in our data set (Table 1), an observation in line with the findings of Imai and Nakai (11). Unfortunately, homology-based inferences became random for the other proteins, dropping the overall average substantially below that for LocTree2 (Table 2). Thus, it would be a very bad idea to annotate an entire proteome only with homology-based inference.

Our new method LocTree3 successfully navigates a path through homology-based and *de novo* prediction of localization (Tables 1-2, Supplementary Tables S5–S7, Section S2). The method is so good that it reaches 18-state overall accuracy (*Q*18, Equation (3)) >95% for half of all the proteins that are most strongly predicted, i.e. have highest reliability (Figure 1). For any new query, users can read off the results whether or not their protein is likely to fall into this top set of '>95%' (RI > 70 for eukaryotes, RI > 80 for bacteria, Figure 1), and whether the prediction comes from a homology search with PSI-BLAST or a *de novo* prediction with LocTree2. For instance, LocTree3 predicts 77% of the entire proteome in human through homology-based inference (a few other highlights from Supplementary Table S8: yeast 68%, *Arabidopsis* 61%, *Caenorhabditis elegans* 47%). However, for yeast only 17% of the predictions originated from direct homology inference, the remainder came from direct experimental annotations (Supplementary Table S8). For human, the corresponding numbers were 30% experimental, 47% through homology inference (Supplementary Table S8). Unfortunately, LocTree2 cannot recover for mistakes made by the homology lookup and all our assessment is based on taking the homology lookup when available. Investigating reasons why homology-based inference was wrong did not give a clear answer (Supplementary Section S3). Due to its high overall performance, reduced prediction time and cached prediction results, LocTree3 web server optimizes well for the handling of large-scale data. Therefore, this web server and its downloadable software should provide an ideal starting point to aid the prediction of protein function through localization predictions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

Thanks to Laszlo Kajan, Manfred Roos and Syeda Tanzeem H. Charu (all TUM) for technical assistance; to Marlena Drabik (TUM) for administrative support. Particular thanks to Yana Bromberg (Rutgers) and Edda Kloppmann (TUM) for many crucially helpful discussions; to Rachel First for preparing the cell figures for the website and to Christina Leslie (Columbia) for providing the initial Profile Kernel. Last, and by no means least, thanks to Ioannis Xenarios (Swiss-Prot, Lausanne), Amos Bairoch (SIB, Geneva), Rolf Apweiler and Alex Bateman (EBI, Hinxton), and their teams for maintaining invaluable databases and to all experimentalists who enabled this work by making their data publicly available.

FUNDING

Alexander von Humboldt Foundation through German Federal Ministry for Education and Research; Ernst Ludwig Ehrlich Studienwerk (to T.G.). Funding for open access charge: German Research Foundation (DFG) and the Technische Universität München within the funding program Open Access Publishing.
Conflict of interest statement. None declared.

REFERENCES

1. Bairoch,A. and Apweiler,R. (1997) The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.*, **25**, 31–36.
2. Simpson,J.C. and Pepperkok,R. (2003) Localizing the proteome. *Genome Biol.*, **4**, 240.
3. Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
4. Koonin,E.V. (2000) Bridging the gap between sequence and function. *Trends Genet.*, **16**, 16.
5. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K., Ben-Hur,A. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods.***10**, 221–227.
6. Lee,D., Redfern,O. and Orengo,C. (2007) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, **8**, 995–1005.
7. Yu,C.S., Chen,Y.C., Lu,C.H. and Hwang,J.K. (2006) Prediction of protein subcellular localization. *Proteins*, **64**, 643–651.
8. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
9. Briesemeister,S., Rahnenfuhrer,J. and Kohlbacher,O. (2010) YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.*, **38**, W497–W502.
10. Yu,N.Y., Wagner,J.R., Laird,M.R., Melli,G., Rey,S., Lo,R., Dao,P., Sahinalp,S.C., Ester,M., Foster,L.J. *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
11. Imai,K. and Nakai,K. (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics*, **10**, 3970–3983.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Goldberg,T., Hamp,T. and Rost,B. (2012) LocTree2 predicts localization for all domains of life. *Bioinformatics*, **28**, i458–i465.
14. Alberts,B.J.A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2002) *Molecular Biology of the Cell*, 4th edn. Garland Science, New York
15. Mooney,C., Cessieux,A., Shields,D.C. and Pollastri,G. (2013) SCL-Epred: a generalised de novo eukaryotic protein subcellular localisation predictor. *Amino Acids*, **45**, 291–299.
16. Kuang,R., Ie,E., Wang,K., Siddiqi,M., Freund,Y. and Leslie,C. (2004) Profile-based string kernels for remote homology detection and motif extraction. Proceedings/IEEE Computational Systems Bioinformatics Conference, CSB. Stanford, California, USA, pp. 152–160.
17. Hamp,T., Goldberg,T. and Rost,B. (2013) Accelerating the original profile kernel. *PLoS One*, **8**, e68459.
18. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Mun Chan,W., Eberhardt,R. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
19. Kajan,L., Yachdav,G., Vicedo,E., Steinegger,M., Mirdita,M., Angermuller,C., Bohm,A., Domke,S., Ertl,J., Mertes,C. *et al.* (2013) Cloud prediction of protein structure and function with PredictProtein for Debian. *BioMed. Res. Int.*, **2013**, 398968.
20. Yachdav,G., Kloppmann,E., Kajan,L., Hecht,M., Goldberg,T., Hamp,T., Hönigschmid,P., Schafferhans,A., Roos,M., Bernhofer,M. *et al.* (2014) PredictProtein – an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.*, doi:10.1093/nar/gku366.
21. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
22. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
23. Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.
24. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
25. Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.
26. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
27. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Supporting online material for: LocTree3 prediction of localization

Tatyana Goldberg, Maximilian Hecht, Tobias Hamp, Timothy Karl, Guy Yachdav, Henrik Nielsen, Burkhard Rost & et al.

Table of Contents for Supporting Online Material

Material

Table S1: Data sets for development and evaluation.

Localization	*Eukaryota*	*Bacteria*	*Archaea*
Chloroplast	133	-	-
Chloroplast membrane	11	-	-
Cytosol	220	179	41
Endoplasmic reticulum	10	-	-
Endoplasmic reticulum membrane	65	-	-
Extra-cellular space	596	82	5
Fimbrium	-	16	-
Golgi apparatus	3	-	-
Golgi apparatus membrane	17	-	-
Mitochondria	140	-	-
Mitochondria membrane	87	-	-
Nucleus	320	-	-
Nucleus membrane	5	-	-
Outer membrane	-	6	-
Plasma membrane	40	144	13
Periplasm	-	52	-
Peroxisome	6	-	-
Peroxisome membrane	2	-	-
Plastid	14	-	-
Vacuole	3	-	-
Vacuole membrane	10	-	-
SUM	1682	479	59

Data: number of proteins per localization class with experimentally determined annotations of a single subcellular localization taken from SWISS-PROT release 2011_04 (1) in at HVAL≤0 (2, 3) sequence-unique sets of eukaryotic, bacterial and archaeal proteins. The data sets were used for development of LocTree3 and its predecessor LocTree2 (4).

Table S2: Homology-based inference from sequence-unique sets

PSI-BLAST E-value threshold[*1]	Q18 - Eukaryota (1682 proteins)	Q6 - Bacteria (479 proteins)
10^{-7}	1±1	0
10^{-5}	2±1	0.4±1
10^{-3}	5±1	2±2
10^{-1}	17±2	9±3
1	27±3	25±5
10	32±3	39±6
100	24±2	33±5
100000	22±2	28±5
Random[*2]	22±2	28±5

Data: 1682 eukaryotic and 479 bacterial sequence-unique proteins with an experimental annotation of a single sub-cellular localization extracted from SWISS-PROT release 2011_04, aligned against themselves.

[1*] PSI-BLAST E-value threshold: defines the E-value (5, 6) threshold for a PSI-BLAST (7) hit, which is different to the query protein, to be considered for performance evaluation

[*2] Random: defines the performance of a random prediction in one of eighteen classes in Eukaryota and six classes in Bacteria, with respect to the data distribution among these classes

Note: Q is the overall prediction accuracy (Eqn. 3, Methods); "±" values refer to standard errors (Eqn. 4, Methods)

Table S3: Data sets for independent/additional testing.

Localization	*New2013_hval0*[*1]		*New2014*[*2]	*Human*[*3]
	Eukaryota	*Bacteria*	*Eukaryota*	*Eukaryota*
Chloroplast	10	-	8	-
Chloroplast membrane	14	-	-	-
Cytosol	43	19	25	965
Endoplasmic reticulum (ER)	1	-	1	41
ER membrane	7	-	-	175
Extra-cellular space	112	20	121	744
Fimbrium	-	1	-	-
Golgi apparatus	2	-	2	15
Golgi apparatus membrane	4	-	-	83
Mitochondrion	13	-	1	290
Mitochondrion membrane	7	-	-	112
Nucleus	43	-	34	1524
Nucleus membrane	-	-	-	7
Outer membrane	-	4	-	-
Periplasm	-	5	-	-
Plasma membrane	9	8	6	1020
Peroxisome	1	-	-	25
Peroxisome membrane	1	-	-	13
Plastid	-	-	-	-
Vacuole	1	-	-	-
Vacuole membrane	5	-	-	2
SUM	273	57	198	5016

Data: number of sequences per localization class in the sets of SWISS-PROT proteins used for the independent/additional testing of LocTree3.

[*1] "New2013_hval0" set: at HVAL≤0 redundancy reduced sets of 273 eukaryotic and 57 bacterial proteins, thus containing no protein pair with >20% pairwise sequence identity over 250 residues aligned. Redundancy reduced set of archaeal proteins was too small (18 proteins) to provide meaningful performance estimates and was thus excluded from the analysis.

[*2] "New2014" set: all eukaryotic proteins added to SWISS-PROT between releases 2013_11 and 2014_03, not redundancy reduced. Because the number of corresponding bacterial proteins was too small (10 proteins), they were excluded from the analysis.

[*3] "Human" set: all proteins with an experimental annotation of exactly one localization class in the SWISS-PROT release 2014_03, not redundancy reduced.

Section S1: LocTree3 assessment on multi-localized proteins

LocTree2 and LocTree3 were developed on proteins from the Swiss-Prot release 2011_04. The number of multi-localized proteins in this release was 48 for bacteria (all annotated with two localization classes) and 4556 for eukaryota (4376 with two localization classes, the others with ≥3). Due to the small number, we dropped bacteria. Reducing redundancy at HVAL≤0 on these 4556 left us with 72 sequence-unique proteins. We applied LocTree3 to these and considered the prediction correct if one of the experimentally observed classes had been predicted. Result: Q18=65±12%; while similar to the performance of LocTree2 on the 1682 cross-validate proteins, it compared less favourable to 80±3% for LocTree3. Why did performance drop on those proteins? Clearly, the random expectation was the opposite, i.e. since we allow one mistake we have a higher random performance: picking one right from 18 is tougher than picking 2 and choosing the best-of-two. In short, our suspicion is that today's double annotations as a whole set are not good enough.

We looked at LocTree3 predictions for five misclassified proteins (i.e. proteins for which none of the experimentally annotated localization classes could be picked by LocTree3) with the highest reliability scores (RIs). One of the five proteins (YG4O_YEAST, RI=38) was an uncharacterized protein while for the remaining four we were able to find the experimental evidence for the predicted localization classes in other sources rather than Swiss-Prot: (1) ZYM1_SCHPO is a metallothionein, which is annotated to be localized to the nucleus and the cytoplasm in SWISS-PROT. LocTree3 predicts this protein to be secreted with the RI=98, we found an experimental evidence for metallothioneins to be secreted in Moltedo *et al.* (8); (2) GPX41_MOUSE is annotated to localize to the mitochondrion and the cytoplasm, while LocTree3 predicts nucleus with the RI=93, which is confirmed by Yant *et al.* (9); (3) NPC2_ASPOR is annotated to be cytoplasmic and a Golgi apparatus protein, LocTree3 however predicts it to be vacuolar with the RI=43, which is true for the protein's ortholog NPC2_YEAST; (4) PEN2_CAEEL is annotated to be localized to the ER membrane and Golgi membrane, LocTree3 predicts mitochondria membrane with RI=36 which is true according to the work of Hansson *et al.* (10). Interestingly, for the protein with the lowest prediction reliability index (CSN4_BRAOL, RI=6) and the predicted localization class chloroplast we could find an evidence in Xiangjun *et al.* (11) stating that the protein acts as a suppressor of chloroplast development. SWISS-PROT annotates the protein to be nuclear and cytoplasmic.

From these findings we conclude that the number of sequence-unique multi-localized proteins as we have them today in SWISS-PROT is rather small and the annotations of multiple localization may be fuzzy and incomplete. Therefore, assessing prediction methods on these proteins may lead to underestimated results and incorrect implications.

Figure S1:

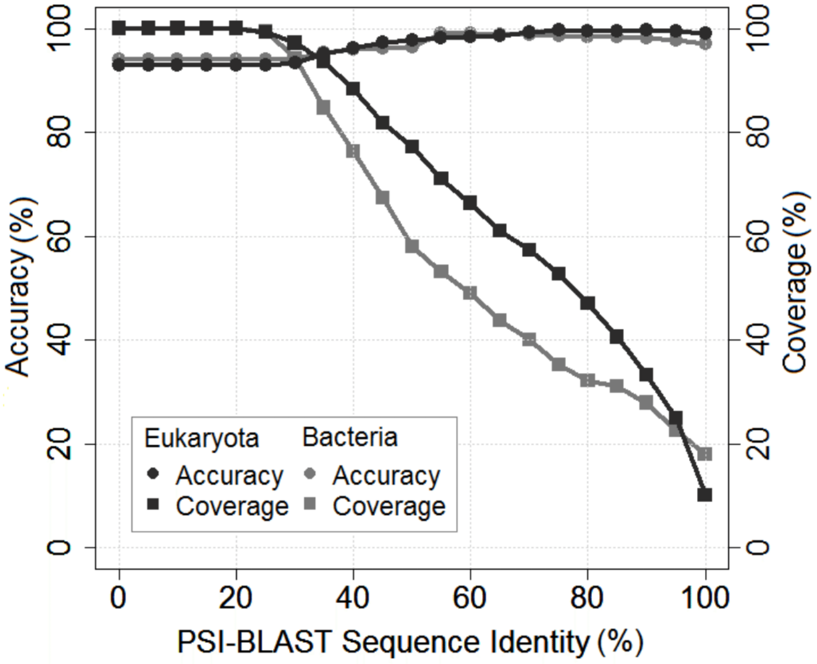


Fig. S1: PSI-BLAST sequence identities to LocTree3 reliability scores. Localization annotation from sequence homologs is more accurate at higher PSI-BLAST pairwise sequence identity (PIDE) values. Here we show the percentage Accuracy/Coverage (Methods) at the given sequence identity thresholds for 995 eukaryotic and 202 bacterial proteins that had a PSI-BLAST hit with E-value≤10^{-3} (6, 7). Since method's performance did not change for PIDE<20, we formed LocTree3's reliability index by normalizing the sequence identity values according to (PIDE-20)*10/8.

Note, the slight decrease of the Accuracy curves at PIDE approaching 100% results from the changed annotations in SWISS-PROT between releases 2011_04 and 2013_11. Though these proteins are predicted to be localized correctly in 2013_11, they are considered as false predictions in the current evaluation (Eukaryota: AIM37_YEAST, ECP_MACFA; Bacteria: ESPR_MYCTU).

Figure S2:

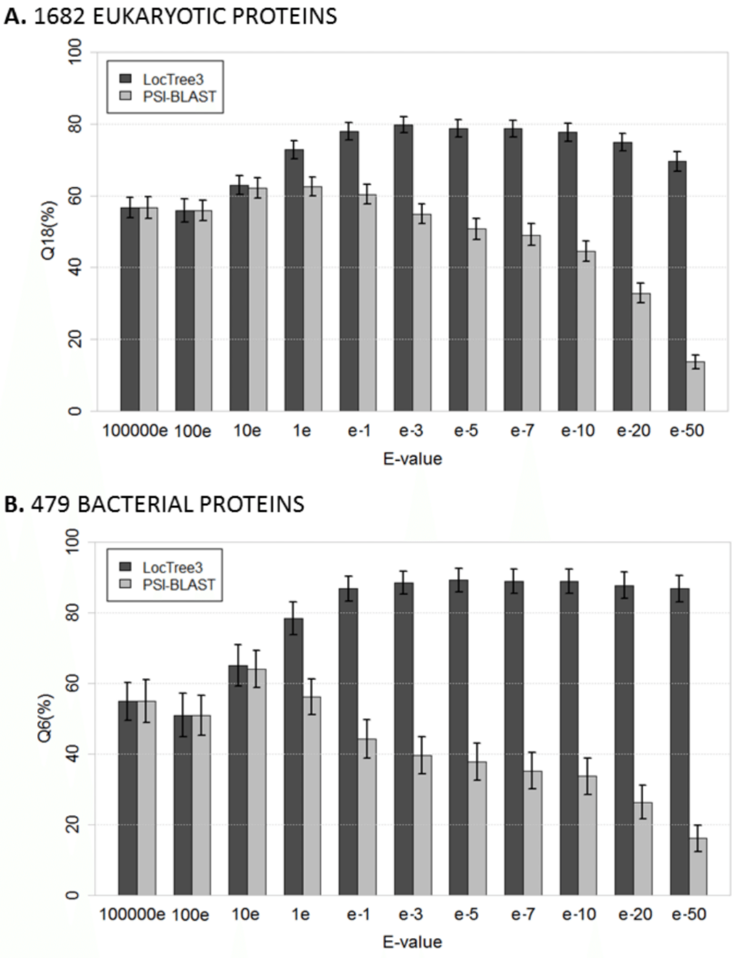


Fig. S2: E-value thresholds for the homology-based inference from all experimentally annotated proteins in SWISS-PROT release 2011_04
The accuracy of localization annotation transfer from sequence homologs (entire SWISS-PROT release 2011_04: 34583 eukaryotic and 4765 bacterial proteins) varies at different PSI-BLAST E-values. Shown is the overall accuracy of LocTree3 (dark grey) and PSI-BLAST (light grey) in predicting 18 localization classes (Q18, Methods) for eukaryotes (Panel A) and 6 classes for bacteria (Panel B) at the given E-value cut-off. PSI-BLAST E-value thresholds reached their peak at high E-value≤10. However, in order to determine the threshold at which value to use LocTree2 and at which PSI-BLAST, we also need to consider the performance of the final merger LocTree3 at the same threshold. The optimal threshold for LocTree3 seemed to be much more conservative, namely at E-value≤10^{-3}.

Table S4: Strategies for annotation transfer by homology.

Method / *Performance*		*Minimum E-val*	*Maximum HVAL*	*Maximum PIDE*	*Majority vote*
Euka-ryota	*Q(18), 1682 proteins*	54 ± 3	53 ± 3	**55 ± 3**	53 ± 3
Bac-teria	*Q(6), 479 proteins*	**40 ± 6**	38 ± 5	**40 ± 5**	39 ± 5

Data: sequence-unique sets of 1682 eukaryotic and 479 bacterial proteins extracted from SWISS-PROT release 2011_04. For each protein a PSI-BLAST profile was built using a combination of UniProt (1) and PDB (12) databases redundancy reduced at 80% sequence identity. The profiles were then aligned at the standard E-value of 10^{-3} (6, 7) against 34583 experimentally annotated eukaryotic and 4765 bacterial proteins available in SWISS-PROT in 2011_04. Given a list of homologs for a query protein we investigated which of the following strategies contributed most to the overall performances Q18 (i.e. correct classification of a protein in one of 18 classes) for Eukaryota and Q6 (i.e. correct classification of a protein in one of 6 classes; Methods) for Bacteria:

Minimum E-val: take the annotation of the hit with the minimum expectation value

Maximum HSSP-val: take the annotation of the hit with the maximum HVAL (9, 10)

Maximum PIDE: take the annotation of the hit with the maximum pairwise sequence identity

Majority vote: take the localization class of most hits

When more than one hit fit the same (e.g. maximum PIDE), we picked the first.

Figure S3:

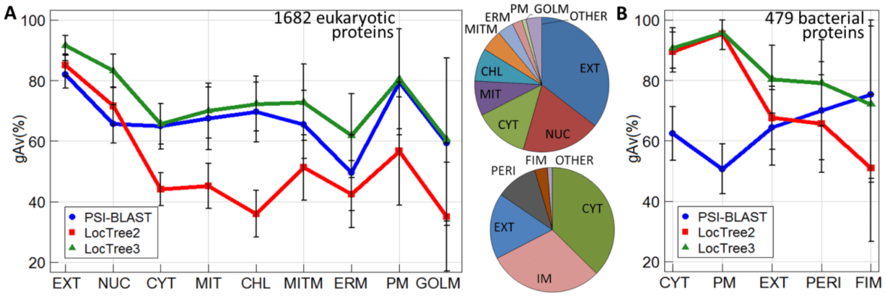


Fig. S3: Class-wise performance comparison of LocTree3 to its sources

PSI-BLAST marks a simple 'lookup' in the database of experimentally annotated proteins from the SWISS-PROT release 2011_04 (i.e. 34583 eukaryotic and 4765 bacterial proteins), self-hits are excluded; LocTree2 is a *de novo* machine learning-based predictor, results shown here are valid for cross-validation on 1682 eukaryotic and 479 bacterial proteins. LocTree3 combines the results of previous two methods by taking PSI-BLAST hits with E-value$\leq 10^{-3}$ and maximum PIDE, if available, and LocTree2 predictions otherwise. We tested on a non-redundant data set of (A) 1682 eukaryotic and (B) 479 bacterial proteins extracted from SWISS-PROT release 2011_04. The localization classes (compartments) on the x-axes mark the averages over all proteins in that class. Note that the x-axes are sorted by the prevalence of that class in the experimental annotations (as given by the inlet pie-charts). In this graph, we force PSI-BLAST to always return a prediction. The y-axes show the geometric average (gAv, Methods) between accuracy and coverage. The pie charts in the centre show the fraction of proteins belonging to each class. LocTree2 predicted classes with most experimental annotations best (A: EXT+NUC, B: CYT+IM+EXT). We could not confirm the same trend for the simple PSI-BLAST protocol. Overall, our new method, LocTree3, published in the web server still maintains a small correlation between performance and experimental annotations with respect to the compartments.

Abbreviations: gAv, geometric average; CHL, chloroplast; CYT, cytosol; ERM, endoplasmic reticulum membrane; EXT, extra-cellular; FIM, fimbrium; GOLM, Golgi apparatus membrane; MIT, mitochondria; MITM, mitochondria membrane; NUC, nucleus; PERI, periplasmic space; PM, plasma membrane.

Table S5: LocTree3 assessment on sequence-unique sets of 479 bacterial and 1682 eukaryotic proteins

Localization	*Nprot*	*Acc*	*Cov*	*gAv*
Extra-cellular	596	88 ± 3	96 ± 2	92 ± 4
Nucleus	320	81 ± 5	86 ± 5	83 ± 6
Cytosol	220	68 ± 7	64 ± 8	66 ± 7
Mitochondria	140	74 ± 10	66 ± 10	70 ± 8
Chloroplast	133	72 ± 9	73 ± 10	72 ± 9
Mitochondria membrane	87	77 ± 11	69 ± 11	73 ± 11
ER membrane	65	67 ± 16	57 ± 14	62 ± 13
Plasma membrane	40	84 ± 15	78 ± 16	81 ± 16
Golgi membrane	17	69 ± 31	53 ± 29	61 ± 27
Plastid	14	50 ± 50	29 ± 31	38 ± 23
Chloroplast membrane	11	80 ± 29*	73 ± 29*	76 ± 32*
ER	10	71 ± 47*	50 ± 35	60 ± 33
Vacuole membrane	10	100*	40 ± 31	63 ± 32
Q(18) – Eukaryota	1682	80 ± 3		
Cytosol	179	91 ± 5	90 ± 5	91 ± 7
Plasma membrane	144	96 ± 4	95 ± 4	96 ± 5
Extra-cellular	82	75 ± 11	87 ± 9	80 ± 11
Periplasm	52	82 ± 14	77 ± 14	79 ± 15
Fimbrium	16	83 ± 25*	63 ± 35	72 ± 26
Q(6) – Bacteria	479	89 ± 4		

Data sets and the LocTree3 performance estimation as in Figure S3. Abbreviations used: *Nprot*, the number of proteins with known localization; *Acc*, accuracy; *Cov*, coverage; *gAv*, geometric coverage of *Acc* and *Cov*; *Q(n)*, overall prediction accuracy. Standard errors were estimated by bootstrapping (Methods).
Note 1: *Q(n)* is a six-state value for bacteria, i.e. the overall accuracy for classification in one of six localization classes, and an eighteen-state value for Eukaryota (Methods). Note 2: Only performances for localization classes containing more than ten proteins are reported.
* = unrealistic upper or lower bound given by the standard error due to the small data set size.

Table S6: Performance comparison on LocTree3's development data

Method		*Eukaryota*				*Bacteria*		
		"Complete" set (1682)[*7]	*¬PSI-BLAST hits (687)*[*7]	*PSI-BLAST hits (995)*[*7]		*"Complete" set (479)*[*8]	*¬PSI-BLAST hits (277)*[*8]	*PSI-BLAST hits (202)*[*8]
Cello 2.5[*1]	Q10	65±3	60±5	70±4	Q5	82±4	81±5	83±6
PSORTb 3.0[*2]		-	-	-		57±5	47±7	71±7
Wolf Psort[*3]		60±3	57±5	63±3		-	-	-
YLoc[*4]		60±3	55±5	64±4		-	-	-
LocTree2[*5]		65±3	62±4	68±4		86±4	86±5	85±6
LocTree3*[6]		**81±3**	**62±4**	**94±2**		**90±3**	**86±5**	**94±4**

[*1] Cello 2.5: employs a system of Support Vector machines to classify eukaryotic proteins in 12 and bacterial in 5 classes using sequence-derived features (13)

[*2] PSORTb 3.0: predicts four classes for Gram-positive and five classes for Gram-negative bacteria through a combination of several classifiers into a Bayesian network (14)

[*3] Wolf Psort: k-nearest neighbour classifier that predicts 12 localization classes for eukaryotes from sequence-derived features (15)

[*4] YLoc: uses sequence-derived features together with GO terms to classify eukaryotic proteins in 11 localization classes through Naïve Bayes (16)

[*5] LocTree2: *de novo* machine learning-based method, results valid for cross-validation

[*6] LocTree3: combines *de-novo* (LocTree2) and homology-based (PSI-BLAST) searches; it uses PSI-BLAST predictions (lookup at E-value≤10^{-3} in a database of experimentally annotated proteins) if available and LocTree2 (results from the cross-validation setting), otherwise

[*7] data set Eukaryota: 1682 sequence-unique eukaryotic proteins in SWISS-PROT release 2011_04; for 995 of those we found PSI-BLAST hits, for 687 we did not

[*8] data set Bacteria: 479 sequence-unique bacterial proteins in SWISS-PROT release 2011_04; for 202 of those we found PSI-BLAST hits, for 227 we did not

Note: Q is the overall prediction accuracy (Eqn. 3, Methods); "±" values refer to standard errors (Eqn. 4, Methods); bold face: "winner in each column"

Table S7: Performance comparison on human protein data

Method	*Q10 (Eqn. 3, Methods)* *"Human proteins" set (5016)*[*6]
Cello 2.5[*1]	75±1
Wolf Psort[*2]	71±1
YLoc[*3]	76±1
LocTree2[*4]	76±1
LocTree3[*5]	**89±1**

[*1-5] Methods as in Table S6

[*5] data set "Human proteins": 5016 human proteins with an experimental annotation of exactly one localization class in SWISS-PROT release 2014_03. A vast majority of these proteins constitutes the training sets of the methods tested.

Note: "±" values refer to standard errors (Eqn. 4, Methods); bold face: "winner in each column"

Table S8: Proteome-wide localization predictions using PSI-BLAST

Organism name	#proteins predicted[*1]	#PSI-BLAST predictions[*2] (% in relation to #proteins predicted)	#Self-hits[*3] (% in relation to #PSI-BLAST predictions)
H. sapiens	20249	15671 (77%)	4638 (30%)
S. cerevisiae	6434	4372 (68%)	2209 (51%)
A. thaliana	27270	16527 (61%)	1843 (11%)
C. elegans	20791	9780 (47%)	346 (4%)
B. weihenstephanensis	5650	1862 (33%)	1 (<1%)
A. pernix	1700	133 (8%)	2 (1%)

[*1] number proteins predicted with LocTree3 in the proteomes of six completely sequenced organisms downloaded from http://www.ebi.ac.uk/genomes/

[*2] number of proteins predicted by PSI-BLAST. The numbers in parenthesis are fractions in relation to the total number proteins predicted in an organism

[*3] number of PSI-BLAST self-hits, i.e. hits that were identical to query proteins. The numbers in parenthesis are fractions in relation to the total number proteins predicted by PSI-BLAST

Section S2: LocTree3 is much more reliable than blind homology-inference. Two recent advances in molecular biology make it impossible to blindly trust annotations. The first are high-throughput experiments that typically change the value of an annotation from, e.g. "protein Q is native in the Golgi" to "protein Q has been detected to have entered the secretory pathway with a probability of 0.7". Clearly, using the second statement to annotate Q as extra-cellular would be very wrong. But what if we added "secretory pathway" as a new "class", should we then annotate it as in that class, or should we maintain the probability? If we maintained the probability: should this be counted as "localization annotated"? What about a protein Q2 that is sequence similar to Q: should we annotate its localization also to be "secretory pathway with 70% chance"? One simple experimental data point generates so many questions that cannot be answered without generating new problems! Thousands of such data points are being created by modern molecular biology every month.

The second problem is contained in the first, but much more prevalent in today's databases that are still heavily based upon detailed biochemical experiments. Assume that we have a reliable annotation for Q as Golgi: how to treat proteins that are related to Q? For instance, those related in terms of sequence similarity. This brings up the argument of Imai & Nakai (17), namely that PSI-BLAST predicts localization more accurately than *de novo* methods. Here we showed that this is true to some extent (Table 1: for some proteins PSI-BLAST is better than LocTree2), but that if predictions are forced, the opposite becomes true (Table 1: averaged over all proteins PSI-BLAST is much less accurate than LocTree2). Clearly, the tool we make available now, LocTree3, settles the discussion. Even if we were right that LocTree3 is the best method currently available to predict protein localization, should we apply it to annotate localization in databases that are exclusively based on experiments such as SWISS-PROT (1)? We suggest a negative answer: leave experimental annotations as clean as possible. Should we then remove almost 90% of (stand Feb. 2014) all annotations about localization in SWISS-PROT (i.e. those based on non-experimental findings)? What about a database that pulls in automated annotations such as UniProt and/or GO (18)? Naïve users querying UniProt might get the impression that over 5m (million) proteins have annotations for localization when the best we can do to develop prediction methods is dig out a list of may be 25k (thousand), i.e. 200 times fewer than suggested by that naïve sieve through UniProt. Clearly, we argue that it would be better to remove the 5m-25k inferred annotations and replace those by LocTree3 predictions marked as predictions and by possibly augmenting this with predictions for all other 45m proteins in today's UniProt (total 52m in Feb. 2014).

Section S3: Possible sources for PSI-BLAST mis-predictions

The idea behind LocTree3 is to use PSI-BLAST if it finds hits and LocTree2, otherwise. Thus if a prediction of the sub-cellular localization is incorrect and is derived from PSI-BLAST, it cannot be 'corrected' by LocTree2 anymore.

Nevertheless, we looked into the cases for which PSI-BLAST annotated proteins incorrectly. In our development eukaryotic data of 1682 eukaryotic proteins, 995 proteins were classified by PSI-BLAST and for remaining 687 proteins it failed to identify a homolog in the data set of all experimentally annotates proteins. Of 995 predicted proteins 69 were misclassified. The most commonly mis-classified pairs of classes (one observed, the other looked up from homolog) were: mitochondria and chloroplast (9 times), plastid and chloroplast (8 times), cytoplasm and nucleus (8 times), cytoplasm and secreted (6 times), cytoplasm and mitochondria (5 times).

These pairs either resembled compartments that are either close in space (e.g. cytoplasm and nucleus), closely related (chloroplasts present one of the three types of plastid) or are very similar in their structure (chloroplast and mitochondria). Therefore, the PSI-BLAST mis-classifications may originate from incorrect experimental annotations, as well as from similarity in translocation signals. About 33% of the mistakes originated from "honest orthologs" (e.g. RK32_EUGGR annotated chloroplast but predicted plastid as its ortholog RK32_ASTLO). The mis-classification with the highest score (PIDE=88%) was made for ECP_MACFA, a protein for which the SWISS-PROT has changed since LocTree3 development from cytosol to be secreted, the latter correctly identified by PSI-BLAST. In other word, this mistake was based on an incorrect earlier annotation.

References for Supporting Online Material

1. A. Bairoch and R. Apweiler, *Nucleic acids research*, **28**, 45-48. (2000)
2. B. Rost, *Protein engineering*, 85-94 (1999)
3. C. Sander, R. Schneider, *Proteins*, 56-68 (1991)
4. T. Goldberg, *et al.*, *Bioinformatics.* **28**, i458-i465 (2012)
5. D. Przybylski and B. Rost, *Proteins*, **46:**197-205 (2002)
6. S.F. Altschul and W. Gish, *Methods in enzymology*, **266:**460-480 (1996)
7. S.F. Altschul, *et al., Nucleic acids research*, **25:**3389-3402 (1997)
8. O. Moltedo, *et al.*, *The Journal of biological chemistry*, **275**, 31819-31825 (2000)
9. L.J. Yant, *et al.*, Free radical biology & medicine, **34**, 496-502 (2003)
10. C.A. Hansson, *et al.*, The Journal of biological chemistry, **279**, 51654-51660 (2004)
11. X. Zhou, *et al.*, BMC plant biology, **11**, 169 (2011)
12. H.M. Berman, *et al., Nucleic acids research*, **28**(1):235-242 (2000)
13. C.S. Yu, *et al., Proteins*, **64**, 643-651 (2006)
14. N.Y. Yu, *et al.*, Bioinformatics, 26, 1608-1615 (2010)
15. P. Horton, *et al.*, Nucleic acids research 2007, 35, W585-W587 (2007)
16. S. Briesemeister, *et al., Bioinformatics* 2010, **26**, 1232-1238 (2010)
17. K. Imai and K. Nakai, *Proteomics* 2010, **10**, 3970-3983 (2010)
18. E.C. Dimmer, *et al., Nucleic acids research* 2012, **40**, D565-570 (2012)

3.3 References

1. Goldberg T, Hamp T, Rost B: **LocTree2 predicts localization for all domains of life.** *Bioinformatics* 2012, **28:**i458-i465.
2. Mooney C, Cessieux A, Shields DC, Pollastri G: **SCL-Epred: a generalised de novo eukaryotic protein subcellular localisation predictor.** *Amino acids* 2013, **45:**291-299.
3. Imai K, Nakai K: **Prediction of subcellular locations of proteins: where to proceed?** *Proteomics* 2010, **10:**3970-3983.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic acids research* 1997, **25:**3389-3402.
5. Hamp T, Goldberg T, Rost B: **Accelerating the Original Profile Kernel.** *PloS one* 2013, **8:**e68459.
6. Kuang R, Ie E, Wang K, Siddiqi M, Freund Y, Leslie C: **Profile-based string kernels for remote homology detection and motif extraction.** *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB IEEE Computational Systems Bioinformatics Conference* 2004**:**152-160.

4 Prediction of nuclear import and nuclear protein sorting

4.1 Introduction

The nucleus is a membranes-enclosed organelle found in eukaryotic cells. It was the first organelle to be discovered as early as in the 17th century [1]. The nucleus contains most of the genetic material, organized into chromosomes, and is also the side for DNA replication and transcription. Nuclear proteins are synthesized in the cytoplasm and their transportation into the nucleus operates differently than to the other sub-cellular compartments, as it occurs through a large structure in the nucleus double membrane, the nuclear pore complex [3]. For this reason, proteins can be transported in their fully folded confirmation. Protein transport can occur between the cytoplasm and the nucleus bi-directionally; often this is done through binding to specific proteins, called karyopherins. Karyopherins bind via recognition of nuclear localization signals (NLS) for protein import or nuclear export signals (NES) for nuclear export in the amino acid sequence of their cargo proteins [4].

Similar to the compartmentalization of a cell, the nucleus is divided into several morphologically distinct compartments, each associated with a different function. However, unlike cellular compartments, nuclear compartments are not membrane-enclosed and are highly dynamic. Studies have shown that nuclear components can be in continuous flux between the compartments and some compartments are formed only during certain cell stages through interaction with proteins, RNA and DNA [5, 6]. It has been suggested that protein translocation within the nucleus also operates through NLS- and NES-like signals [7, 8]. However, this mechanism is not well understood [5].

In this Chapter, novel method, LocNuclei, is described that associates nuclear proteins with 13 sub-nuclear compartments at a high level of overall accuracy Q13 = 62%. Similar to LocTree3 (described in Chapter 3), this is done through the combination of homology information to proteins with known sub-nuclear association and machine learning predictions. In addition, the method is able to predict if a nuclear protein is functional in other sub-cellular compartments (*e.g.* the mitochondria) at the level of overall accuracy Q2 = 72%. Applied to 6,230 human proteins, predicted to localize to the nucleus, we identified 77% of them to be functional at the nucleoplasm (30% of all annotations), chromatin (17%), nucleolus (17%) or PML bodies (13). Plugging in protein-protein interaction data, we found most intra-nuclear interactions to occur between proteins of these four localizations.

4.2 Compartmentalization of the nucleus

In this work, we distinguished between 13 sub-nuclear compartments, shown in Figure 1 and briefly described below.

Cajal bodies: Cajal bodies are small structures that contain coiled threads of the marker protein, coilin. The interaction of coilin with other proteins within the Cajal bodies appears to increase their functional efficiency, *e.g.* the modification and assembly of splicing machinery [9]. The number and size of the Cajal bodies varies between tissues and organisms, as well as during different differentiation and development stages [10]. Generally, Cajal bodies are found in cells of high transcriptional activity and splicing demands, such as in neuronal and cancer cells [11].

Chromatin: Chromatins are fibrous structures forming the chromosomes. The major proteins of the chromatin are histones, whose function has been shown to be mediated through post-translational modifications [12]. The chromatin is formed through binding of histones to the DNA. Its functions include the regulation of gene expression, DNA replication and segregation during cell division, as well as DNA damage recognition and repair [13]. Chromatin-associated proteins contain a high diversity of motifs, many of which are specific to protein-protein interactions. Thus, chromatin proteins appear to abundantly interact among each other and with other nuclear proteins, *e.g.* proteins involved in transcription and replication [14].

Nuclear envelope: The nuclear envelope is a barrier that separates the contents of the nucleus from the cytoplasm, and regulates the trafficking of proteins and other molecules between these two compartments. In addition, nuclear envelope serves as an anchoring site for chromatin that correctly positions the chromosomes within the nucleus, and for the cytoskeleton that correctly positions the nucleus within the cell [15]. The nuclear envelope consists of two membranes, the outer and the inner nuclear membranes, which are like other cellular membranes are phospholipid bilayers [12]. The membranes are continuous with the Endoplasmic Reticulum (ER), though each of them is associated with proteins that are not enriched in the ER [15].

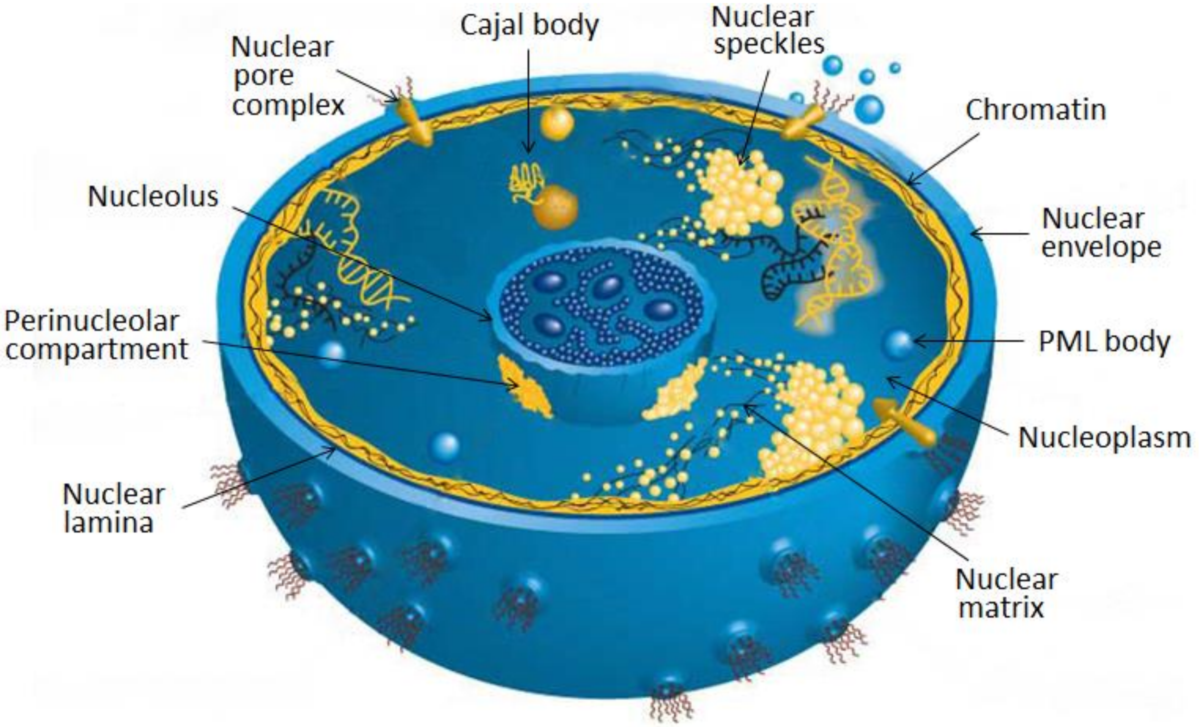


Figure 1: Sub-nuclear compartments. The figure shows sub-nuclear compartments predicted by LocNuclei (the spindle apparatus and the kinetochores are not shown). Figure adapted from [16].

Nuclear lamina: The nuclear lamina is an essential component of metazoan cells, but is not found in unicellular organisms and plants [17] . It is localized at the interface between chromatin and the inner nuclear membrane. The nuclear lamina is composed of lamins, type V intermediate filament proteins, and many lamins-interacting proteins. This layer was found to be also closely associated with the nuclear pore complexes [17]. The functions of the nuclear lamina include DNA replication, RNA transcription, chromatin organization, cell cycle regulation, cell development and differentiation, nuclear migration, and apoptosis [18, 19]

Nuclear matrix: The nuclear matrix is a network of fibrous structure extending throughout the whole interior of the nucleus [20]. The exact function role if the nuclear matrix remains unclear [21] . Though, proteins that were found to be associated with the matrix are known to be involved in a number of nuclear activities, such as DNA replication and transcription, and RNA processing and transport [22].

Nuclear pore complex: A nuclear pore complex is a highly structured assembly of proteins that form a tunnel across the nuclear envelope for the regulated transport of proteins and other molecules across it. A single nuclear pore complex comprises up to 500 copies of approximately 50 highly conserved distinct proteins [15, 23]. The nucleus of the human cell can contain up to 3,000 nuclear pore complexes [23]. In addition to the main function as the nuclear gatekeepers, the proteins of the nuclear pore complexes were found to be associated with other functions, such as the regulation of gene expression [23].

Nuclear speckles: Nuclear speckles are dynamic structures located in the interchromatin region of the nucleoplasm and enriched in pre-mRNA splicing factors. They serve as the side of storage, assembly and modification of splicing factors. The splicing event however does not occur at nuclear speckles. The size and the number of nuclear speckles vary between different cell types and within a cell type. Furthermore, the components of nuclear speckles can be exchanged with the nucleoplasm and other nuclear compartments [24]. Some of the speckle components exhibit a speckle targeting signal [14].

Nucleolus: The nucleolus is the largest, densest and the best studied sub-nuclear compartment. There are one or several nucleoli in mot eukaryotic cells [25]. Although nucleoli are most famous for the functional role in the biogenesis of ribosomes, they are also involved in numerous other processes, including RNA modification, cell-cycle control and stress response [26]. Furthermore, the nucleolus constraints the movement of chromatin, which implicates its role in higher-order chromatin arrangement [27]. Over 700 distinct proteins have been associated to localize in the nucleolus [28] and a nucleolar targeting signal has recently been described [29].

Nucleoplasm: The nucleoplasm is, similarly to the cytoplasm, a highly gelatinous liquid that is held within the nuclear envelope and that acts as a suspension substance for nuclear compartments. It is rich in protein enzymes and other material required for the synthesis of RNA and DNA [30, 31]. Another major constituent of the nucleoplasm are chromosomes. The nucleoplasm plays an important role in the maintenance of the nuclear shape and the transport of molecules between the nucleus and the cytoplasm.

Perinucleolar compartment: The perinucleolar compartment is an irregularly shaped structure found at the periphery of the nucleolus. This compartment is largely found in transformed and cancer cells [32, 33]. It forms at late telophase and disassembles at the beginning of mitosis [32]. The perinucleolar compartment in enriched with by RNA polymerase III transcribed RNAs and RNA-binding proteins, many of which are exchanged with other sub-nuclear compartments [33].

PML bodies: PML bodies are dynamic nuclear matrix-associated structures that require the ProMyelocytic Leukaemia (PML) protein for their formation and incorporate a number of other proteins that shuffle between the PML bodies and other sub-nuclear compartments [34]. PML bodies play a role in transcription, apoptosis promotion, post-translational modifications, suppression of oncogenic transformation, DNA repair and antiviral defense [15, 34].

Kinetochore: The kinetochores are multiprotein control modules that anchor segregating chromosomes to spindle microtubules and enforce their correct movement to two opposite poles of the spindle apparatus. The kinetochores are assembled during cell division (mitosis and meiosis), and many of their components are highly dynamic and cycle between the kinetochores and the spindle apparatus [34].

Spindle apparatus: The spindle apparatus segregates chromosomes during cell division in two daughter cells. The spindle apparatus is organized by centrosomes and constitutes spindle poles, kinetochores and hundreds of microtubule-associated proteins [35]. The apparatus is located at two opposite poles of the cell to ensure the separation of replicating chromosomes in two exactly equal sets. The failure of correct chromosomes segregation can lead to chromosomal instability, aneuploidy or tetraploidy (both leading to cancer) and cell death [35].

4.3 Materials and Methods

Data sets for development and evaluation

We downloaded experimentally annotated nuclear proteins together with their annotations of the sub-nuclear localization, if available, from the HPRD [36], NMPdb [37], NOPdb [38], NPD [39], NSort/DB [40] and Swiss-Prot [41] databases. Because databases use different terms for annotations of some sub-nuclear compartments, we normalized annotations from different databased to a set of fixed keywords, *e.g.* we normalized 'PML-NBs' and 'Nuclear dots' to 'PML bodies' (Supplementary Table S1 in the Appendix). This resulted in a set of 13 distinct keywords describing our sub-nuclear data set.

Out of total 12,055 proteins annotated experimentally as nuclear, only 3,522 (29%) proteins were additionally annotated to be associated with one or more sub-nuclear compartment. We homology reduced this set at HVAL < 20 [42, 43] using UniqueProt [44]. For alignments longer than 250 residues, HVAL < 20 implies a maximal pairwise sequence identity of 40% [43]. At lower HVALs, the data set became too small for meaningful performance estimates (*e.g.* at HVAL < 0, we had in five of 13 classes less than 10 proteins annotated to be localized in the corresponding class). The final sequence unique sub-nuclear set comprised 1,934 proteins (Table S2).

Furthermore, out of total 12,055 nuclear proteins, 4,722 were annotated to be localized in at least one additional sub-cellular compartment (*e.g.* the mitochondria). We homology reduced this set of 12,055 proteins at HVAL < 0 (maximal pairwise sequence identity of 20% over 250 residues aligned) and obtained 1,098 sequence-unique proteins, of which 559 were annotated to exclusively localize to the nucleus.

The resulted prediction method was thus trained to differentiate between (i) proteins localized to the nucleus and proteins localized to the nucleus and other sub-cellular compartments, as well as between (ii) proteins of 13 sub-nuclear localization classes.

Prediction methods

Similarly to LocTree3 [45], a high performance method for the prediction of protein sub-cellular localization (Chapter 3), LocNuclei combines homology-based predictions if available with *de novo* predictions otherwise. We determined all parameters for our final predictor LocNuclei in a five-fold cross-validation setting, *i.e.* we split the entire Development set into five equally-sized subsets. We trained five models, each on a different combination of four of these subsets, and tested them on the remaining one.

Homology-based predictions: We transferred annotations by homology using PSI-BLAST [46] alignments. For all proteins of known localization, we generated PSI-BLAST profiles with two iterations and E-value ≤ 10^{-3} using an 80% non-redundant database combining UniProt [47] and PDB [48]. We then aligned these profiles at E-value ≤ 10^{-3} against non-redundant proteins in our Development set (1,888 proteins for the prediction of 13 sub-nuclear localizations and 1,037 proteins for the prediction of nuclear proteins shuffling to other sub-cellular compartments). For performance estimates, we excluded the PSI-BLAST self-hits. Similar to LocTree3, we transferred annotation to the query protein from the hit with the highest pairwise sequence identity of all retrieved alignments.

De novo prediction: We used the Support Vector Machine (SVM) [49] implementation of LibSVM [50] and the Profile Kernel function [51, 52] (Chapter 2). We trained 13 different SVM classifiers to predict 13 sub-nuclear localizations, where each classifier was trained to discriminate between proteins of a particular sub-nuclear class and proteins of all other classes. To predict nuclear proteins that are travelers to other sub-cellular compartments, we separately trained another SVM.

NSort: NSort [53] is a framework, developed in 2010, of eight Bayesian Network-based classifiers that predict protein sub-nuclear localization in eight classes (nucleolus, perinucleolar region, PML bodies, nuclear speckle, Cajal bodies, chromatin and nuclear pore complexes). Each classifier operates from biological features including protein sequence, proteins interactions, domain and post-translational modification. Each prediction of NSort can be traced back to the feature contributing most to the result.

4.4 Results and Discussion

High performance values: Q13 = 62% and Q2 = 72%

LocNuclei is a predictor developed to discriminate between (i) proteins of 13 sub-nuclear localization classes and (ii) proteins localized to the nucleus only and proteins localized to the nucleus and other sub-cellular compartments. For each of these two prediction tasks, we developed our predictor in a five-fold cross-validation setting using the sequence-unique Development set (Chapter 4.2) and optimized the parameters of its components (PSI-BLAST and SVM-based inferences) separately.

For the prediction task of 13 sub-nuclear compartments, the homology-based inference for proteins for which experimentally annotated homologs were available achieved the highest level of overall performance (Chapter 2.2) Q13 = 68% at E-value ≤ 10^{-50} (Figure 2, black arrow). However, when applied to the complete set, the performance at the same E-value dropped significantly to Q13 = 18%. This results was still significantly above random (<8%), showing that the homology-based inference works, though the annotations of sub-nuclear localizations are largely missing. On the same test set, our *de novo*-based inference employing a battery of 13 SVM classifiers achieved an almost 3-fold higher level of Q13 = 59%. This result encouraged us to use a simple protocol, introduced in our previous work, LocTree3 [45], that unites PSI-BLAST whenever possible and the SVM if no PSI-BLAST results were available. We chose the PSI-BLAST E-value of 10^{-20} as the decision threshold between PSI-BLAST and *de novo* inferences. The combined method, LocNuclei, outperformed both its components, reaching an overall accuracy Q13 = 62 ± 3% (Figure 2).

Similarly, for the second prediction task, we combined homology-based PSI-BLAST with the Profile Kernel SVM to predict nuclear proteins functional in other sub-cellular compartments. We found LocNuclei to perform best at the PSI-BLAST E-value ≤ 10^{-5}, reaching an overall performance Q2 = 72 ± 2% (Figure S1 in the Appendix).

LocNuclei best on novel proteins

Comparing prediction performance of our method to the published performance of NSort (the only available sub-nuclear predictor during the development of LocNuclei) has only little value due to differences in the training and test sets. Running NSort on our independent sets (*i.e.* proteins experimentally annotated after the development of NSort) was also problematic, because NSort's source code was no longer available. Thus, the only

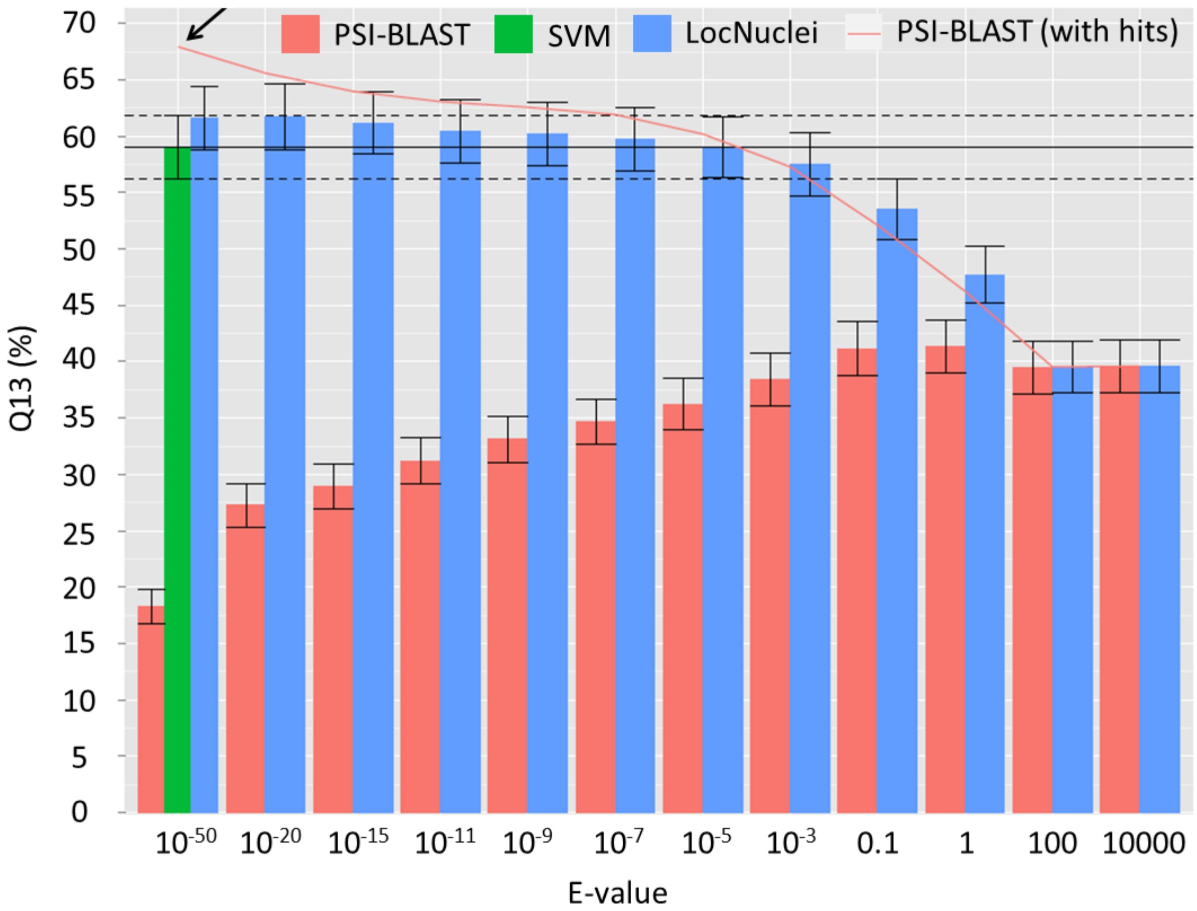


Figure 2: E-value thresholds for the homology-based component of LocNuclei (prediction of 13 sub-nuclear classes). The accuracy Q13 (Chapter 2.2) for classifying proteins in 13 sub-nuclear compartments using the homology-based inference with PSI-BLAST (based on 3,522 experimentally annotated proteins) varies at different E-value thresholds. For proteins for which a homolog is available, the highest accuracy Q13 = 68% is achieved at E-value ≤ 10^{-50} (black arrow). However, if considering proteins for which no homology is available, this value drops to 18%. The performance of SVMs on the same set is Q13 = 59% (black horizontal line, dotted lines mark the values considering the standard error). To determine, at which E-value threshold to use PSI-BLAST and at which the ensemble of 13 SVMs, we needed to consider the performance of the final method LocNuclei at the same threshold. We found LocNuclei to be most conservative at E-value ≤ 10^{-20}.

meaningful way to benchmark the performance of these two methods was to train and test LocNuclei on the exactly same set as NSort was trained and tested upon. Towards this end, we downloaded the data set of LocNuclei from http://nsort.org/db/ and split it into five subsets to train our model on four of them and to test on the remaining one. We rotated these sets five times, so that each protein in the NSort set was tested exactly once. We computed area under the ROC curve (AUC) from the average of five splits as the performance estimate. For training, we used the parameters that we found to perform best

Sub-nuclear compartment	*Number of proteins*	*AUC NSort*	*AUC LocNuclei*
Perinucleolar	24	0.80	0.82
Cajal body	49	0.60	0.72
Nuclear pore complex	51	0.79	0.91
Nuclear lamina	77	0.70	0.83
PML bodies	91	0.77	0.81
Chromatin	323	0.71	0.80
Nuclear speckle	403	0.71	0.79
Nucleolus	598	0.60	0.74
Sum/ Mean	**1,285**	**0.71**	**0.80**

Table 1: Performance comparison of LocNuclei to NSort. We used the development data of NSort, comprising 1,285 sequence-unique proteins annotated with eight sub-nuclear localization classes to train LocNuclei. For training, we used those parameters that we found to perform best on our development set. On proteins from all eight sub-nuclear localizations tested, LocNuclei outperformed NSort. The overall cross-validated prediction accuracy of LocNuclei was 0.80, while that of NSort was 0.71. The values for NSort were extracted from the corresponding publication [53].

on LocNuclei's development set. The data set of NSort contained proteins of eight sub-nuclear localizations; for all of them LocNuclei outperformed NSort (Table 1). The mean AUC (over all eight compartments) was 0.71 for NSort and 0.80 for LocNuclei. Thus, we could show that the improvement of LocNuclei originated from the underlying method advancement and not from the difference in the composition of the data sets.

Over 30% of nuclear proteins predicted to reside in the nucleolus

After completing the development, we applied LocNuclei to the human nucleosome protein data. Towards this end, we downloaded the reference human proteome from the European Bioinformatics Institute (EBI: http://www.ebi.ac.uk/reference_proteomes) and identified nuclear and nuclear membrane proteins in it using LocTree3 [45]. The resulted data set of 6,230 proteins was then provided to LocNuclei as input. We predicted over 1/3 of them to be travelers, *i.e.* localize to additional sub-cellular compartments other than nucleus. For about 11% of all nuclear proteins we could not predict any sub-nuclear localization, for 36%

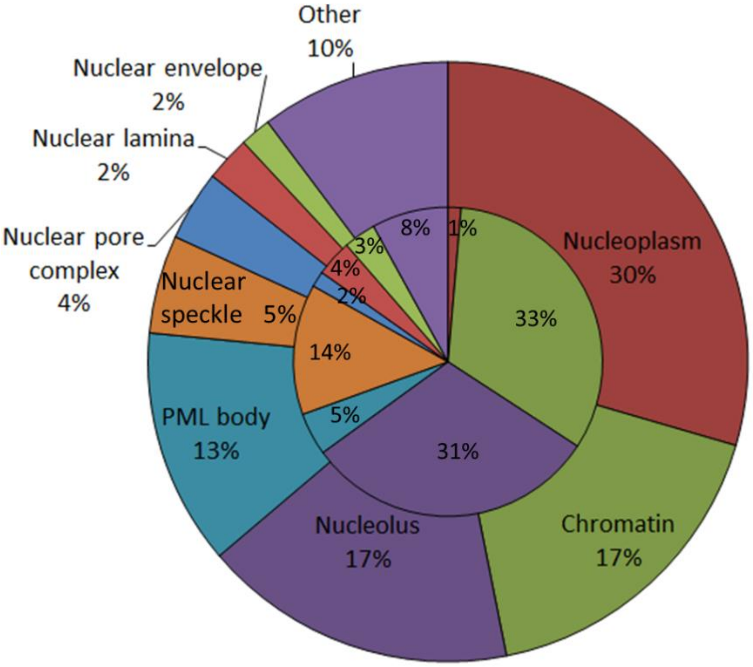


Figure 3: Composition of sub-nuclear compartments in the human proteome and LocNuclei's development set. The *inner ring* of the pie chart represents the composition of LocNuclei's development set (assembled from nuclear proteins of various organisms), while the *outer ring* represents the composition of human 6,230 nuclear proteins (predicted by LocTree3 [45] from EBI's human reference proteome). Both data sets differ significantly in their composition. Thus, the composition of the predicted human nucleosome is not just a reflection of the development set.

proteins we predicted localization in one sub-nuclear compartment, and for remaining 53% localization in at least two compartments. Furthermore, we predicted 30% of all proteins to be associated with the nucleoplasm (Figure 3), which is a large aquatic compartment surrounding the nucleus interior. Many proteins including enzymes, specific receptors of hormones and of other effectors, proteins of yet unknown function, as well as proteins shuttling between the nucleus and the cytoplasm are found in the nucleoplasm [30, 31]. The second largest predicted sub-nuclear localization compartment was chromatin (17%), a structure that is built from the interaction with the DNA. The role of the chromatin is in the maintenance of DNA and the regulation of its transcription. It is known that many proteins that compose the chromatin are exchanged with other sub-nuclear compartments, such as the nucleolus [14, 54], which is the third largest class of human nuclear proteins (17%) predicted by LocNuclei. Overall, the composition of predicted sub-nuclear compartments in human did not resemble that of our development set, suggesting that there is no correlation between the predictions of both sets and the predicted compartmentalization of the human nucleosome is likely to reflect its true composition.

Most protein-protein interactions take place between four sub-nuclear compartments

Protein-protein interactions (PPIs) are central to almost all biological processes. Thus, to better understand biological mechanisms, the knowledge of PPIs that underlie them is indispensable [55]. We used the set of human proteins with the predicted annotations of sub-nuclear localizations and mapped them to the experimentally determined protein-protein interaction data from the Human Protein Reference Database (HPRD) [36]. We found most protein-protein interactions to occur within and between four largest sub-nuclear compartments, *i.e.* nucleoplasm, chromatin, nucleolus and PML bodies (Figure 4). Furthermore, the proteins of nuclear speckles appeared to abundantly interact with the

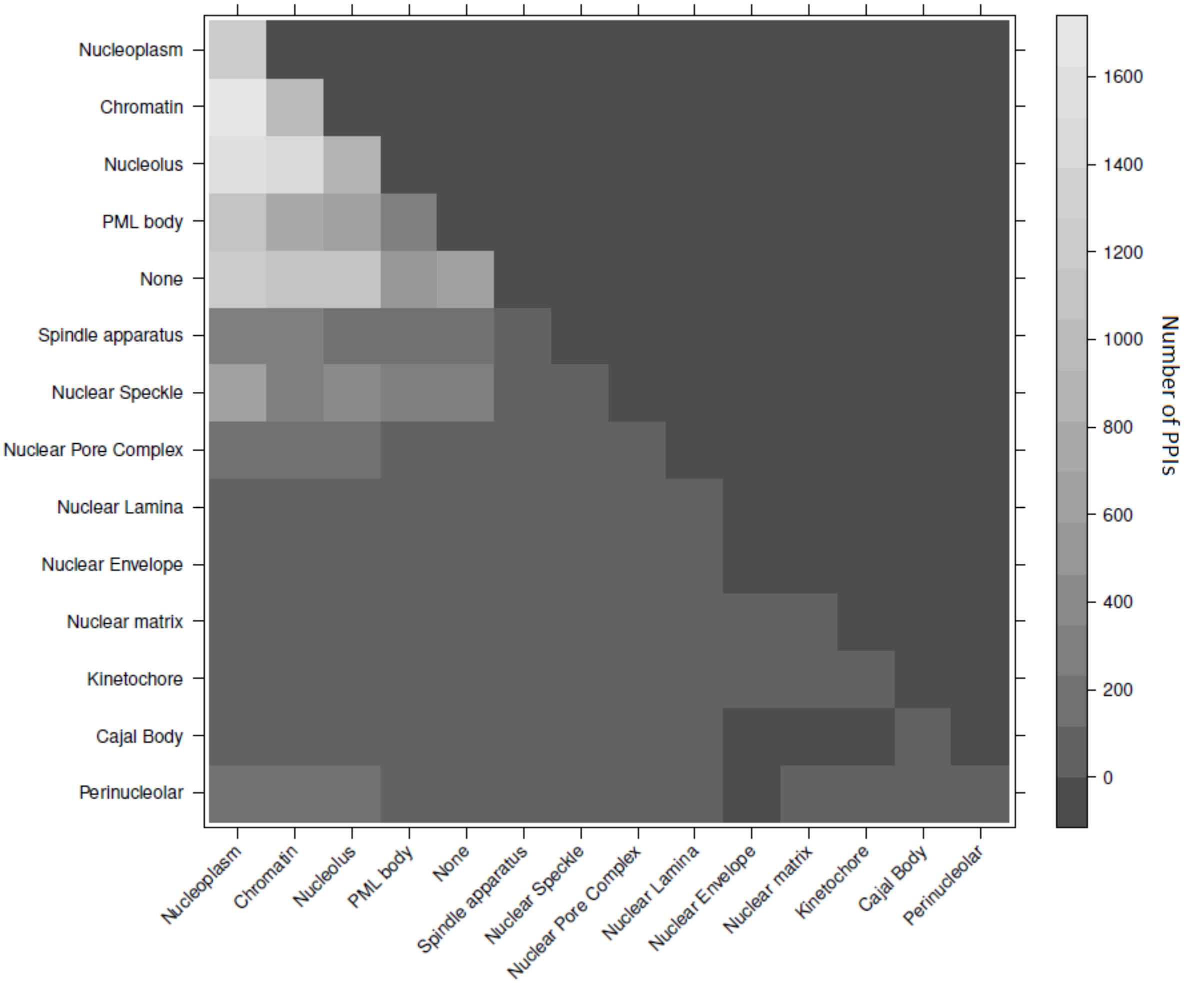


Figure 4: Number of protein-protein interactions (PPIs) in the human nucleosome. The figure plots the number of PPIs within and between 13 sub-nuclear compartments. We extracted experimentally annotated PPIs from the HPRD database [36] and mapped those in human proteins of predicted 13 sub-nuclear compartments. PPIs are most frequent within four largest compartments (nucleoplasm, chromatin, nucleolus and PML bodies; light gray colored cells) and between them.

proteins localized at nucleoplasm and nucleolus. Interestingly, we identified perinucleolar proteins, which compose the smallest class of by LocNuclei predicted sub-nuclear compartments (<0.4% of all annotations in the human nucleosome), to be another outlier in the high number of PPIs. Namely, we identified perinucleolar proteins to often bind proteins of the neighboring compartments: nucleolus, nucleoplasm and chromatin.

Nuclear proteins tend to be disordered

Recent studies have shown that in order to function, some proteins may not adopt unique three dimensional structures in isolation [56]. Instead, functional proteins may contain largely unstructured regions (30 amino acids and more) that sample a large portion of their available conformational space. These proteins are called *disordered*. Studies of different genomes have shown that disorder is very abundant in nature and can be more frequently observed in eukaryotes than in other domains of life [56-60]. Furthermore, it has been shown that many disordered proteins are nuclear [61], involved in *e.g.* DNA and RNA binding [62, 63], nuclear pore transport [64] and transcription [65]. In this experiment, we analyzed the prevalence of protein disorder within the nucleus and compared it to other sub-cellular compartments. We predicted protein disorder with NorsNet [66], a machine learning-based method that predicts unstructured regions of 70 or more consecutive residues.

Using NorsNet on human proteins with by LocTree3 predicted sub-cellular localization annotations, we identified disordered proteins to be over five times more frequent in the nucleus than in mitochondria (mean: 55% vs. 10%; Figure 5A) and almost twice as frequent in the nucleus than in the extra-cellular space (mean: 55% vs. 31%). The distribution of disordered proteins within the individual sub-nuclear compartments also varied substantially (Figure 5B). We identified strongest preferences for disordered proteins at the compartments of dynamic structures, such as nuclear matrix (mean: 98%) and nuclear speckles (mean: 86%). The lowest percentage of disordered proteins was identified at rather stable complexes, such as nuclear lamina, nuclear envelope, kinetochores and the nuclear pore complexes (all below 15%; Figure 5B).

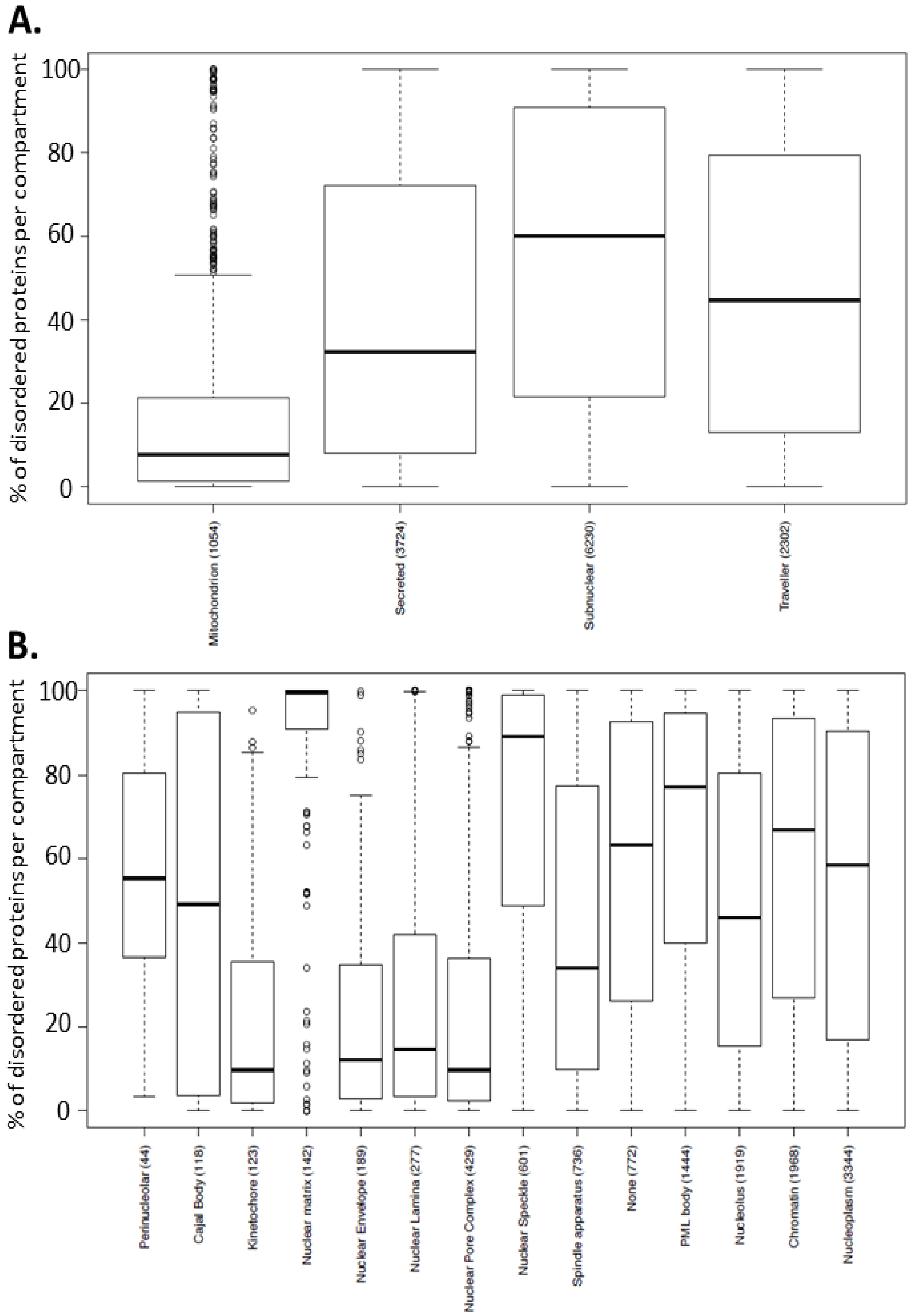


Figure 5: Distribution of disordered proteins in the nucleus and other sub-cellular compartments. We predicted protein disorder (at least 70 unstructured consecutive residues) using NorsNet [66] and plotted the fraction of disordered proteins (Y-axis) in different sub-cellular and sub-nuclear compartments (X-axis). Numbers in parenthesis are numbers of proteins annotated to localize in a particular compartment. **(A)** The highest fraction of disordered proteins appears to be in the nucleus (mean: 55%), compared to mitochondrial (10%) and secreted (31%) proteins. Note, nuclear proteins with additional localizations in other sub-cellular compartments are less disordered (travelers: 44%) than the sum of all nuclear proteins (55%). **(B)** Within the nucleus, the most disordered are nuclear matrix and nuclear speckle proteins, while the least disordered are proteins localized at nuclear pore complexes, kinetochores, nuclear envelope and nuclear lamina.

4.5 References

1. Erhardt M, Adamska I, Franco OL: **Plant nuclear proteomics--inside the cell maestro.** *The FEBS journal* 2010, **277:**3295-3307.
2. Bell EA, Boehnke P, Harrison TM, Mao WL: **Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon.** *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112:**14518-14521.
3. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell, 4th edition.* New York: Garland Science; 2002.
4. Freitas N, Cunha C: **Mechanisms and signals for the nuclear import of proteins.** *Current genomics* 2009, **10:**550-557.
5. Carmo-Fonseca M: **The contribution of nuclear compartmentalization to gene regulation.** *Cell* 2002, **108:**513-521.
6. Chubb JR, Bickmore WA: **Considering nuclear compartmentalization in the light of nuclear dynamics.** *Cell* 2003, **112:**403-406.
7. Lohrum MA, Ashcroft M, Kubbutat MH, Vousden KH: **Identification of a cryptic nucleolar-localization signal in MDM2.** *Nature cell biology* 2000, **2:**179-181.
8. Eilbracht J, Schmidt-Zachmann MS: **Identification of a sequence element directing a protein to nuclear speckles.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98:**3849-3854.
9. Morris GE: **The Cajal body.** *Biochimica et biophysica acta* 2008, **1783:**2108-2115.
10. Ogg SC, Lamond AI: **Cajal bodies and coilin--moving towards function.** *The Journal of cell biology* 2002, **159:**17-21.
11. Hebert MD: **Signals controlling Cajal body assembly and function.** *The international journal of biochemistry & cell biology* 2013, **45:**1314-1317.
12. Cooper JM: *The Cell: A Molecular Approach, 2nd edition.* Boston University: Sunderland (MA): Sinauer Associates; 2000.
13. Dimitrova E, Turberfield AH, Klose RJ: **Histone demethylases in chromatin biology and beyond.** *EMBO reports* 2015.
14. Bickmore WA, Sutherland HG: **Addressing protein localization within the nucleus.** *The EMBO journal* 2002, **21:**1248-1254.
15. Hetzer MW, Walther TC, Mattaj IW: **Pushing the envelope: structure, function, and dynamics of the nuclear periphery.** *Annual review of cell and developmental biology* 2005, **21:**347-380.
16. Hooper C: **The nucleus and sub-nuclear domains.** *Abcam discover more* 2006.
17. Dechat T, Adam SA, Taimen P, Shimi T, Goldman RD: **Nuclear lamins.** *Cold Spring Harbor perspectives in biology* 2010, **2:**a000547.
18. Gruenbaum Y, Goldman RD, Meyuhas R, Mills E, Margalit A, Fridkin A, Dayani Y, Prokocimer M, Enosh A: **The nuclear lamina and its functions in the nucleus.** *International review of cytology* 2003, **226:**1-62.
19. Gruenbaum Y, Margalit A, Goldman RD, Shumaker DK, Wilson KL: **The nuclear lamina comes of age.** *Nature reviews Molecular cell biology* 2005, **6:**21-31.
20. Verheijen R, van Venrooij W, Ramaekers F: **The nuclear matrix: structure and composition.** *Journal of cell science* 1988, **90 (Pt 1):**11-36.
21. Pederson T: **Half a century of "the nuclear matrix".** *Molecular biology of the cell* 2000, **11:**799-805.

22. Stuurman N, Meijne AM, van der Pol AJ, de Jong L, van Driel R, van Renswoude J: **The nuclear matrix from cells of different origin. Evidence for a common set of matrix proteins.** *The Journal of biological chemistry* 1990, **265:**5460-5465.
23. Kabachinski G, Schwartz TU: **The nuclear pore complex--structure and function at a glance.** *Journal of cell science* 2015, **128:**423-429.
24. Spector DL, Lamond AI: **Nuclear speckles.** *Cold Spring Harbor perspectives in biology* 2011, **3**.
25. Shaw PJ, Jordan EG: **The nucleolus.** *Annual review of cell and developmental biology* 1995, **11:**93-121.
26. Raska I, Shaw PJ, Cmarko D: **Structure and function of the nucleolus in the spotlight.** *Current opinion in cell biology* 2006, **18:**325-334.
27. Nemeth A, Conesa A, Santoyo-Lopez J, Medina I, Montaner D, Peterfia B, Solovei I, Cremer T, Dopazo J, Langst G: **Initial genomics of the human nucleolus.** *PLoS genetics* 2010, **6:**e1000889.
28. Andersen JS, Lam YW, Leung AK, Ong SE, Lyon CE, Lamond AI, Mann M: **Nucleolar proteome dynamics.** *Nature* 2005, **433:**77-83.
29. Martin RM, Ter-Avetisyan G, Herce HD, Ludwig AK, Lattig-Tunnemann G, Cardoso MC: **Principles of protein targeting to the nucleolus.** *Nucleus* 2015, **6:**314-325.
30. Cameron I: *Acidic Proteins of the Nucleus.* Elsevier; 2012.
31. Plopper G: *Principles of Cell Biology.* Rensselaer Polytechnic Institute George Plopper: Jones & Bartlett Publishers; 2014.
32. Huang S, Deerinck TJ, Ellisman MH, Spector DL: **The perinucleolar compartment and transcription.** *The Journal of cell biology* 1998, **143:**35-47.
33. Pollock C, Huang S: **The perinucleolar compartment.** *Journal of cellular biochemistry* 2009, **107:**189-193.
34. Everett RD, Chelbi-Alix MK: **PML and PML nuclear bodies: implications in antiviral defence.** *Biochimie* 2007, **89:**819-830.
35. Schmidt S, Essmann F, Cirstea IC, Kuck F, Thakur HC, Singh M, Kletke A, Janicke RU, Wiek C, Hanenberg H, et al: **The centrosome and mitotic spindle apparatus in cancer and senescence.** *Cell Cycle* 2010, **9:**4469-4473.
36. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al: **Human Protein Reference Database--2009 update.** *Nucleic acids research* 2009, **37:**D767-772.
37. Mika S, Rost B: **NMPdb: Database of Nuclear Matrix Proteins.** *Nucleic acids research* 2005, **33:**D160-163.
38. Leung AK, Trinkle-Mulcahy L, Lam YW, Andersen JS, Mann M, Lamond AI: **NOPdb: Nucleolar Proteome Database.** *Nucleic acids research* 2006, **34:**D218-220.
39. Dellaire G, Farrall R, Bickmore WA: **The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome.** *Nucleic acids research* 2003, **31:**328-330.
40. Willadsen K, Mohamad N, Boden M: **NSort/DB: an intranuclear compartment protein database.** *Genomics, proteomics & bioinformatics* 2012, **10:**226-229.
41. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic acids research* 2000, **28:**45-48.
42. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9:**56-68.
43. Rost B: **Twilight zone of protein sequence alignments.** *Protein engineering* 1999, **12:**85-94.
44. Mika S, Rost B: **UniqueProt: Creating representative protein sequence sets.** *Nucleic acids research* 2003, **31:**3789-3791.

45. Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, Altermann U, Angerer P, Ansorge S, Balasz K, et al: **LocTree3 prediction of localization.** *Nucleic acids research* 2014, **42:**W350-355.

46. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic acids research* 1997, **25:**3389-3402.

47. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL.** *Nucleic acids research* 1997, **25:**31-36.

48. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic acids research* 2000, **28:**235-242.

49. Cortes C, Vapnik V: **Support-Vector Networks.** *Machine Learning* 1995, **20:**273-297.

50. Chang CC, Lin CJ: **LIBSVM : a library for support vector machines. .** *ACM Transactions on Intelligent Systems and Technology,* 2011, **2:**27:21--27:27.

51. Kuang R, Ie E, Wang K, Siddiqi M, Freund Y, Leslie C: **Profile-based string kernels for remote homology detection and motif extraction.** *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB IEEE Computational Systems Bioinformatics Conference* 2004**:**152-160.

52. Hamp T, Goldberg T, Rost B: **Accelerating the Original Profile Kernel.** *PloS one* 2013, **8:**e68459.

53. Bauer DC, Willadsen K, Buske FA, Le Cao KA, Bailey TL, Dellaire G, Boden M: **Sorting the nuclear proteome.** *Bioinformatics* 2011, **27:**i7-14.

54. McKeown PC, Shaw PJ: **Chromatin: linking structure and function in the nucleolus.** *Chromosoma* 2009, **118:**11-23.

55. Ofran Y, Rost B: **Protein-protein interaction hotspots carved into sequences.** *PLoS computational biology* 2007, **3:**e119.

56. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B: **Protein disorder--a breakthrough invention of evolution?** *Current opinion in structural biology* 2011, **21:**412-418.

57. Xue B, Dunker AK, Uversky VN: **Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life.** *Journal of biomolecular structure & dynamics* 2012, **30:**137-149.

58. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L, et al: **D(2)P(2): database of disordered protein predictions.** *Nucleic acids research* 2013, **41:**D508-516.

59. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, Hu G, Uversky VN, Kurgan L: **Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.** *Cellular and molecular life sciences : CMLS* 2015, **72:**137-151.

60. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ: **Intrinsic protein disorder in complete genomes.** *Genome informatics Workshop on Genome Informatics* 2000, **11:**161-171.

61. Skupien-Rabian B, Jankowska U, Swiderska B, Lukasiewicz S, Ryszawy D, Dziedzicka-Wasylewska M, Kedracka-Krok S: **Proteomic and bioinformatic analysis of a nuclear intrinsically disordered proteome.** *Journal of proteomics* 2016, **130:**76-84.

62. Homma K, Fukuchi S, Nishikawa K, Sakamoto S, Sugawara H: **Intrinsically disordered regions have specific functions in mitochondrial and nuclear proteins.** *Molecular bioSystems* 2012, **8:**247-255.

63. Peng Z, Kurgan L: **High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.** *Nucleic acids research* 2015, **43:**e121.

64. Toretsky JA, Wright PE: **Assemblages: functional units formed by cellular phase separation.** *The Journal of cell biology* 2014, **206:**579-588.

65. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK: **Intrinsic disorder in transcription factors.** *Biochemistry* 2006, **45:**6873-6888.

66. Schlessinger A, Liu J, Rost B: **Natively unstructured loops differ from other loops.** *PLoS computational biology* 2007, **3:**e140.

5 NLSDb2.0: a database of nuclear import and export signals

5.1 Introduction

Eukaryotic cells transport proteins in and out of the nucleus through nuclear pore complexes. This transport is often mediated by specific molecules, called karyopherins, that recognize nuclear localization signals (NLS) or nuclear export signals (NES) in their cargo proteins [1]. The best experimentally described NLS are monopartite and bipartite signals [2-4]. Monopartite signals are characterized by a short stretch of amino acids, which are mostly basic, and bipartite signals are composed of two monopartite signals separated by a variable 10-12 amino acid linker region [5]. A more recently observed signal is the Proline-Tyrosine NLS (PY-NLS) [6]. PY-NLS can be classified as hydrophobic or basic, dependent on its N-terminal region that is followed by the consensus sequence of an arginine (R), lysine (K) or histidine (H), then a proline and tyrosine (R/H/KX-$_{(2\text{-}5)}$-PY). The classical NES are represented by leucine-reach NES, first identified in HIV-1 [7, 8]. Several solutions have been proposed to describe the consensus sequence of NES [9-11], but they did not suffice to identify new NES-containing proteins [12]. Note that not all nuclear proteins are transported via the signals described above [13, 14]. Furthermore, sequences of many non-nuclear proteins match the sequences of nuclear import and export signals.

NLSdb is a comprehensive database that for the first time attempted to collect all experimentally verified NLS in a single resource [15]. It contains 114 NLS published before 2000. In addition, the database provides amino acid sequences of 194 potential NLS, discovered through "*in silico* mutagenesis" from the set of experimental signals [16]. Several of these potential NLS have already been confirmed experimentally (*e.g.* [17-22]).

In this Chapter, an update of NLSdb to the current state of available data is described. We collected experimentally verified 2,391 NLS published in literature after 2000, and 817 experimentally verified and published NES. We applied the procedure of "*in silico* mutagenesis" [16] to these sets and discovered novel 4,310 potential NLS and 1,768 potential NES. Our final set matched 43% and 28**%** of all known nuclear proteins with NLS and NES, respectively, and none currently known non-nuclear protein. By clustering the sequences of experimental signals, we identified a clear separation of NLS in 40 distinct clusters and of NES in 27 clusters. Thus, the consensus sequence describing each of these clusters can be used as a consensus for a different type of a nuclear signal. NLSdb 2.0 is available online at https://rostlab.org/services/nlsdb2/.

5.2 Materials and Methods

Collection of a trusted set of nuclear signals

We extracted amino acid sequences of experimentally annotated NLS and NES from literature [5, 6, 11, 23-27] and the ValidNES [28], NESbase [10], NESdb [29] and Swiss-Prot (release 2015_01) [30] databases. For literature searches we used the criteria described by the authors of NLSdb [26]. Namely, to accept a signal as experimentally confirmed, the signal needs to be proven sufficient to mediate the nuclear transport of a non-nuclear protein to the nucleus and its deletion must result in the prevention of protein nuclear transport. For Swiss-Prot searches we used keywords 'importin binding signal', '*in vitro* NLS', 'nuclear localization signal', 'bipartite NLS', 'PY-NLS', 'nuclear import signal' and 'signal for nuclear transport' to identify NLS. Accordingly, we used keywords 'nuclear export signal' and 'nuclear export sequence' to identify NES. We included only those annotations that were supported with the following Evidence Codes Ontology (ECO) [31] codes: (i) ECO:0000269 (manually curated information for which there is published experimental evidence); (ii) ECO:0000250 (manually curated information which has been propagated from a related experimentally characterized protein); (iii) ECO:0000305 (manually curated information which has been inferred by a curator based on his/her scientific knowledge or on the scientific content of an article); and (iv) ECO:0000255 (manual assertions for information which has been generated by the UniProt automatic annotation system or by various sequence analysis programs). Signal sequences from other databases were included only if their annotations were supported by experimental findings. Table 1 provides an overview of the number of nuclear signal sequences extracted from each source.

Source	*Lange et al.[23]*	*Lee et al. [6] & Suel et al. [25]*	*SeqNLS [24]*	*NLSdb [16]*	*Swiss-Prot [30]*
Number of unique NLS	104	19	69	114	2,227
Source	*García-Santisteban [27]*	*NESbase [10]*	*VALidNES [28]*	*NESdb [29]*	*Swiss-Prot [30]*
Number of unique NES	32	73	261	175	433

Table 1: Composition of the initial set of NLSdb 2.0 signals. The numbers of unique sequences of nuclear localization signals (NLS) and nuclear export signals (NLS) extracted from each source are provided. Total numbers of unique signals extracted from all sources: 2,391 for NLS and 817 for NES.

Sets of nuclear and non-nuclear proteins

We downloaded protein sequences with specific annotations of sub-cellular localization from Swiss-Prot (release 2015_01). Included were only experimental annotations: (i) tagged with the ECO code ECO:0000269 (manually curated information for which there is published experimental evidence) and (ii) annotations lacking any ECO code and also lacking keywords ‘potential, ‘probable’ or ‘by similarity’, denoting non-experimental evidence. Based on the localization annotation, we sorted proteins in two sets: (i) nuclear proteins (true positives; 6,538 proteins) and (ii) non-nuclear proteins (true negatives; 23,028 proteins). We applied UniqueProt [32] at HSSP-value ≤ 0 [33, 34] to each of these sets individually and identified 761 distinct structural families for nuclear proteins and 2,434 distinct structural families for non-nuclear proteins. We used sets of nuclear and non-nuclear proteins to test the validity of all potential signals, obtained through the “*in silico* mutagenesis” approach (described below). We required sequences of valid NLS and NES to match in nuclear proteins and to not match in non-nuclear proteins.

In silico mutagenesis

To increase the set of trusted (*i.e.* experimentally annotated or by experts verified) NLS and NES by potential, experimentally yet un-identified signals, we applied the “*in silico* mutagenesis” approach, similar to that described in [16]. We performed the following steps:

(i) Starting from the trusted set of 2,391 NLS and 817 NES, we removed signals matching sequences of any 23,028 non-nuclear proteins extracted from Swiss-Prot. The resulted trusted set of signals comprised 310 NLS and 166 NES.

(ii) Then, we changed amino acids at each position of each signal in the reduced trusted set (19 substitutions for each amino acid) and mapped new signals in the sequences of nuclear and non-nuclear proteins. We kept only those potential NLS and NES that matched nuclear proteins and no non-nuclear protein.

(iii) Finally, we shortened each signal from our potential set by one amino acid at each end of the sequence and repeated step (ii). For each potential signal we kept only its shortest sub-sequence matching exclusively in nuclear proteins. The final set of potential signals comprised 4,310 NLS and 1,768 NES.

Signal clustering

To analyze sequence variability of nuclear signals in our sets, we performed the following steps to each of our trusted sets of NLS and NES separately:

(i) *Construct an evolutionary distance matrix for sequences of all signals.* We aligned all-against-all sequences of nuclear signals in our set using the maximum likelihood method, described by Thorne *et al*. [35]. The JTT matrix from the work of Jones *et al.* [36] was used as a rate matrix.

(ii) *Derive a phylogenetic tree from the distance matrix*. We used the “Neighbor” implementation from the PHYLIP package [37] to apply the UPGMA clustering method [38] on the distance matrix to calculate the evolutionary tree.

(iii) *Determine sub-groups within the tree*. To identify distinct subgroups within the tree, or clusters of sequence-similar nuclear signals, we applied a graph-pruning method suggested by Krause *et al.* [39]. Briefly summarized, starting from each leaf until reaching the root, the method calculates at each node of the tree the ratio between the size of the tree of a parent node and the size of the tree at the current note. The node at which the ratio is largest is used as a cut-point and its subtree as a distinct cluster. Following this approach, each leaf (*i.e.* nuclear signal) is assigned to exactly one cluster.

(iv) *Calculate a consensus sequence for each cluster*. For each cluster identified in the previous step, we represented its consensus sequence as a position weight matrix (PWM), generated by aligning all sequences of a cluster using MAFFT [40]. We visualized PWMs as sequence logos [41] using WebLogo [42].

5.3 Results and Discussion

Known signals vary in length and protein sequences they occur in

Our trusted data set of unique nuclear signals contained 1,960 monopartite nuclear localization signals (NLS; 61.1% of the whole data set), 413 bipartite NLS (12.9%), 18 PY-NLS (0.6%) and 817 nuclear export signals (NES; 25.4%). The length distribution of these signals is shown in Figure 1. About one third of signals in our set was formed through monopartite NLS of length ranging between 4 and 10 amino acids (60.2% of all monopartite NLS can be found in this range; Supplementary Figure S2 in the Appendix). Interestingly, the second largest group of monopartite signals (20.1%) falls into the range between 15 and 19 amino acids, where the most of bipartite NLS (61%; Figure S2) can be found. This result suggests possible annotation mistakes of monopartite NLS in this range. Possibly, all monopartite NLS outside the range of the first peak (*i.e.* longer than 15 amino acids, 30% of all monopartite NLS) are in fact bipartite signals. All PY-NLS in our set were between 16 and 36 amino acids long. Finally, typical NES seemed to have a sequence length varying between 9 and 13 amino acids.

Further, we tested whether protein sequences containing similar nuclear signals also tend have a high overall sequence similarity. The monopartite NLS in our trusted set were annotated to localize in sequences of 4,243 unique proteins, the bipartite NLS in sequences of 808 unique proteins, the PY-NLS in sequences of 19 unique proteins and the NES in sequences of 1,715 unique proteins. In total, 3,208 unique signal sequences in our trusted set were annotated in 4,492 unique proteins, indicating that, on average, each signal occurred in more than one protein (the ratio between the number of proteins and the number of signals occurring in these proteins was 1.40; Table 2). We applied cd-hit [43] to the protein set to reduce it at 100%, 80%, 60% and 40% sequence identity and UniqueProt [32] to eliminate all proteins with a pairwise sequence identity over 20%. At 100% sequence identity, we had 4,120 unique proteins with annotations of 3,138 unique nuclear signals (ratio 1.31), implying that at this high sequence identity, the prediction of NLS and NES from sequences of annotated homologous proteins is in principle possible. This is, however, different at lower sequence identity levels. Namely, at sequence identity of 80%, proteins do not have annotations of exactly the same nuclear targeting signals anymore. Thus, the prediction of nuclear signals from sequence homology at levels below 80% sequence identity is likely to fail.

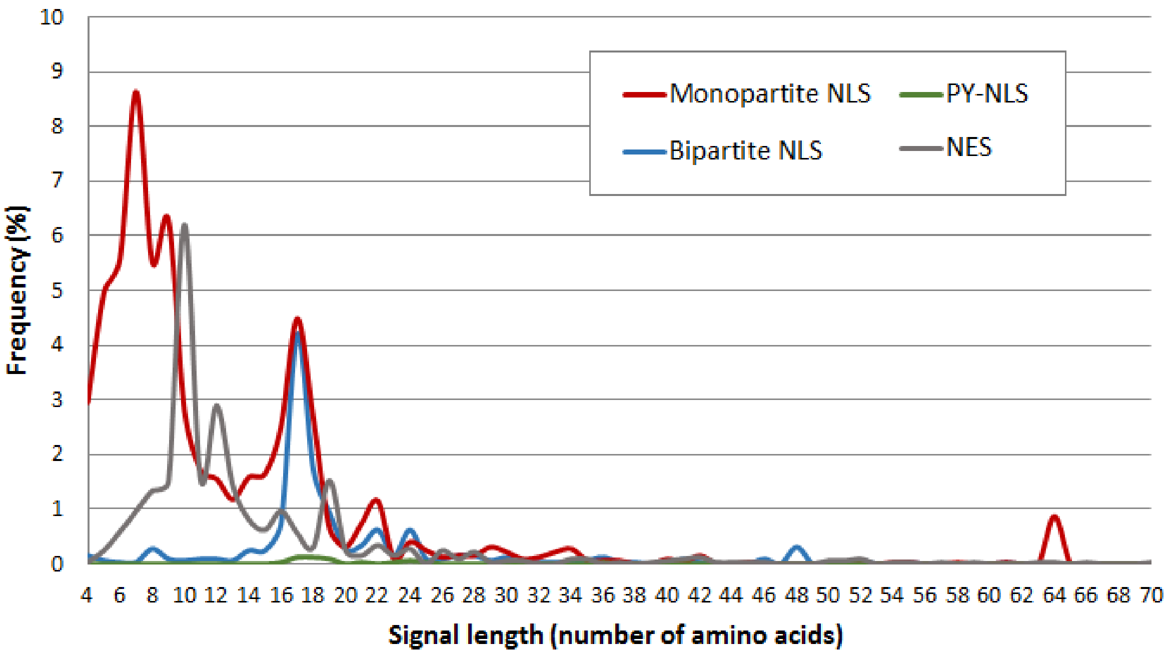


Figure 1: Sequence length distribution of nuclear signals in the trusted set. The trusted set comprised unique 1,960 monopartite NLS, 413 bipartite NLS, 18 PY-NLS and 817 NES, which together formed 100% of the data shown in the figure. Most monopartite signals peak at 4-10 and 15-19 amino acids, with the latter being also the peak for bipartite signals. Probably, monopartite signals with sequence length exceeding 13 residues are erroneously annotated bipartite signals. Note, we do not show results for signals longer than 70 amino acids, as they constitute <1.2% of our trusted set.

. *Sequence identity*	*Number of unique proteins*	*Number of unique nuclear signals*	*Ratio: number of proteins/ number of signals*
All proteins	4,492	3,208	1.40
100%	4,120	3,138	1.31
80%	1,968	2,375	0.82
60%	1,552	1,948	0.79
40%	1,280	1,667	0.77
20%	216	291	0.74

Table 2: Different proteins contain different nuclear signals. The data set of 4,492 distinct proteins containing 3,208 distinct nuclear signals from our trusted set was homology reduced at 100%, 80%, 60% and 40% sequence identity using cd-hit [32] and at 20% sequence identity using UniqueProt [36]. For each redundancy reduced protein set, we monitored the number of annotated nuclear signals in the set. While proteins of high sequence similarity tend to share some of nuclear signals, this is not the case already at similarity levels ≤80%.

Most annotated nuclear signals are human

Nuclear targeting signals in our trusted set were annotated in a high number of 486 distinct species (Table 3). Of all signals, 70% were of virus origin and 29% were eukaryotic. Only monopartite NLS were annotated in bacteria. The high diversity within the virus domain clearly shows the focus of virus-oriented biological research, which is of significant importance for public health [44, 45]. During the infection process, a virus requires host cell's resources to replicate. Most DNA and RNA viruses use nuclear proteins for this process [46-49]. Therefore, the viral genome must enter the nucleus of the host cell. This can only be done using the host nuclear protein transport machinery [47, 50, 51], which is often activated through the recognition of nuclear targeting signals (NLS and/or NES).

Despite the high diversity of species annotations in our trusted set of nuclear signals, the vast majority of them were annotated in sequences of only few species. Figure 2 shows top twelve most frequent species annotations for over 69% of trusted monopartite NLS in our set. We mapped annotations of other signal types in these species and found them to cover over 62% of trusted bipartite NLS, 100% of trusted PY-NLS and over 70% of trusted NES (the distribution of most frequent species annotations for each signal type individually is shown in Figure S3 in the Appendix). Most frequent annotations for all signal types, except PY-NLS (comprising only 19 signals in our set), were made in *Influenza A virus*, which is of all infectious viruses one of the leading causes of death worldwide [52, 53]. The other viruses within the top twelve species were *Hepatitis C virus*, affecting primarily the liver of over 30 million patients alone in the United States [54], and *Human immunodeficiency virus* type 1 (HIV-1) of group M subtype B, the dominant HIV subtype in the Americas,

Domain of organism	*Monopartite NLS (401 species)*	*Bipartite NLS (153 species)*	*PY-NLS (2 species)*	*NES (151 species)*	*All signals (486 species)*
Virus	290 (72.3%)	86 (56.2%)	-	97 (64.2%)	341 (70.1%)
Eukaryota	*107 (26.7%)*	*67 (43.8%)*	*2 (100%)*	*54 (35.6%)*	*141 (29.1%)*
Bacteria	4 (1%)	-	-	-	4 (0.8%)

Table 3: Species annotations in our trusted set of nuclear signals. The numbers of unique sequences of nuclear localization signals (NLS) and nuclear export signals (NLS) extracted from each source are provided. The total number of unique sequences extracted from all sources was 2,391 for NLS and 817 for NES. For all signal types, most annotations were done in viruses. PY-NLS held annotations only in Eukaryota. Monopartite NLS were the only signals with annotations in Bacteria.

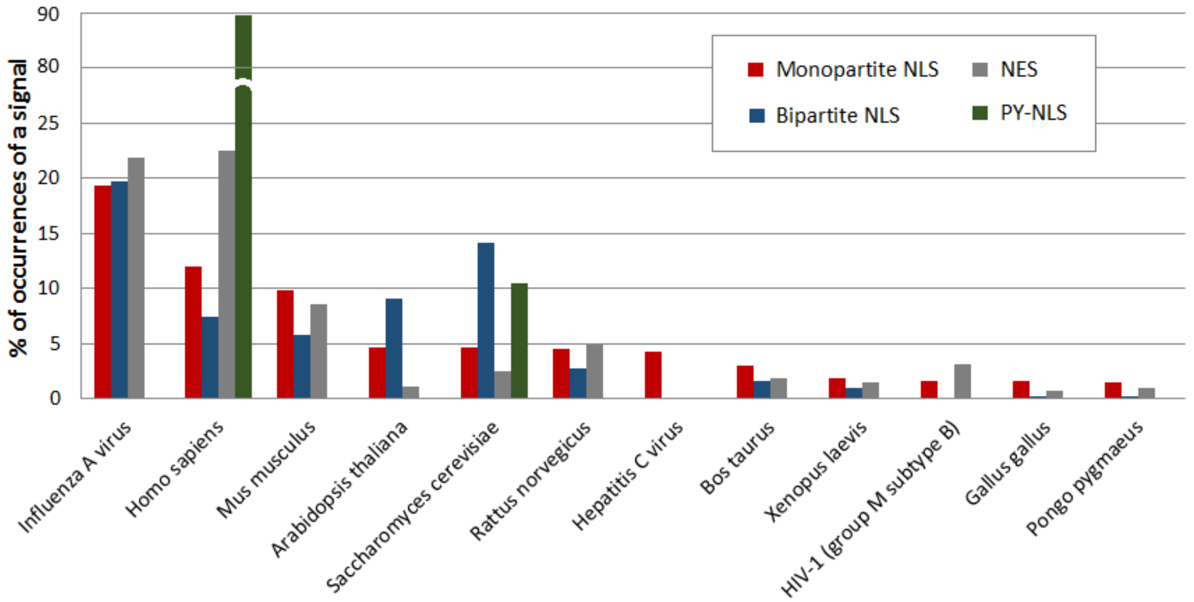


Figure 2: Top twelve most frequent organism annotations of trusted NLS and NES. Over 50% of monopartite NLS, bipartite NLS, and NES, as well as 100% of PY-NLAS are most frequently annotated in proteins of twelve model organisms shown in this figure.

Western Europe and Australasia [55], causing a progressive failure of the immune system of infected patients and a subsequent increased life-threatening risk of opportunistic diseases and cancer [56-58].

In eukaryotes, most annotations of monopartite NLS, PY-NLS and NES were done in human and other model organisms (Figure 2). Interestingly, bipartite NLS annotations mostly came from yeast *S. cerevisiae* (14% of all bipartite NLS annotations) and plants *A. thaliana* and *O. sativa* subsp. Japonica (15.6%; Figure S3B). It is possible that this observation was due to the fact that most research on bipartite signals has so far been done in yeast and plant organisms. However, it is also possible that bipartite signals are more frequent in yeast and plants than in other organisms. To test the second hypothesis we analyzed the distribution of bipartite and monopartite signals (of which one third are likely to be annotation mistakes of bipartite signals; Figure 1) in human, Arabidopsis and yeast (Figure S4). For all organisms, the length distributions formed two clear peaks, between 6 and 9 amino acids (typical range for monopartite NLS; Figure 1), and between 16 and 19 amino acids (typical range for bipartite NLS; Figure 1). Though, the frequencies of these peaks were different. Bipartite NLS appeared indeed to be most frequent in yeast and monopartite NLS most frequent in Arabidopsis.

Nuclear signals form many different clusters of similar sequences

We grouped sequences in each set of monopartite NLS, bipartite NLS, PY-NLS and NES from our trusted set to identify clusters of similar sequences. To describe briefly, this was done by all-against-all aligning sequences from each set separately. Based on the alignment results we built four phylogenetic trees (one tree for each signal type). We identified clusters within these trees, aligned sequences within each cluster and visualized the results as sequence logos [41].

The phylogenetic tree for 1,960 monopartite NLS divided signals in two clusters, a major cluster (“Major”) forming 39 sub-clusters and one cluster (“Minor”) containing sequences of 13 NLS (Figure S5 in the Appendix). In contrast to the common definition of monopartite NLS being a stretch of basic amino acids, amino acids forming the “Minor” cluster appear to be largely acidic and hydrophobic, as shown by the sequence logo in Figure 3. This was different for signals of the “Major” cluster. Figure 4 displays examples of sequence logos of nine randomly chosen its sub-clusters. The logos largely display stretches of highly conserved basic amino acids, though there are also exceptions. For example, Cluster II shows a conserved pattern of basic amino acids that on the N-terminus is preceding by a strongly conserved hydrophobic proline and on the C-terminus is following by three variable residues and a conserved hydrophobic leucine. Similarly, in Cluster III, the core of basic amino acids is preceded by hydrophobic residues and is followed by highly conserved asparagine and valine.

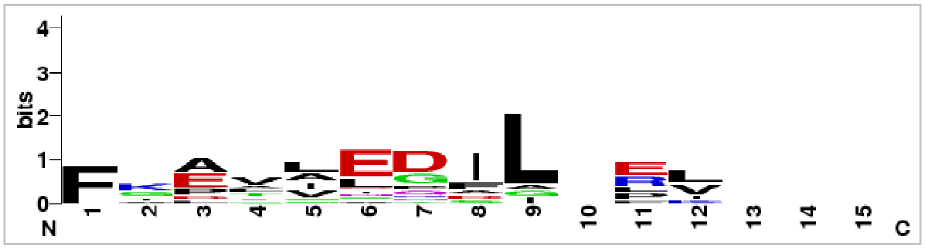


Figure 3: Sequence logo representation of the “Minor” monopartite NLS cluster. Amino acid sequences of 13 monopartite NLS deviate, in contrast to all other monopartite NLS sequences, from the standard description of a stretch of basic amino acids. Thus, these 13 sequences form a separate (“Minor”) cluster in the phylogenetic tree of 1,960 unique monopartite NLS. The sequence logo was generated using WebLogo [42]. Amino acids are colored according to their chemical properties: polar amino acids (G,S,T,Y,C,Q,N) are green, basic (K,R,H) blue, acidic (D,E) red and hydrophobic (A,V,L,I,P,W,F,M) amino acids are black. At each position, amino acids are represented from most frequent (placed on top of a letter stack) to least (placed at bottom). The letter conservation is given by bits (Y-axis) with 4.32 bits being the maximum possible conservation.

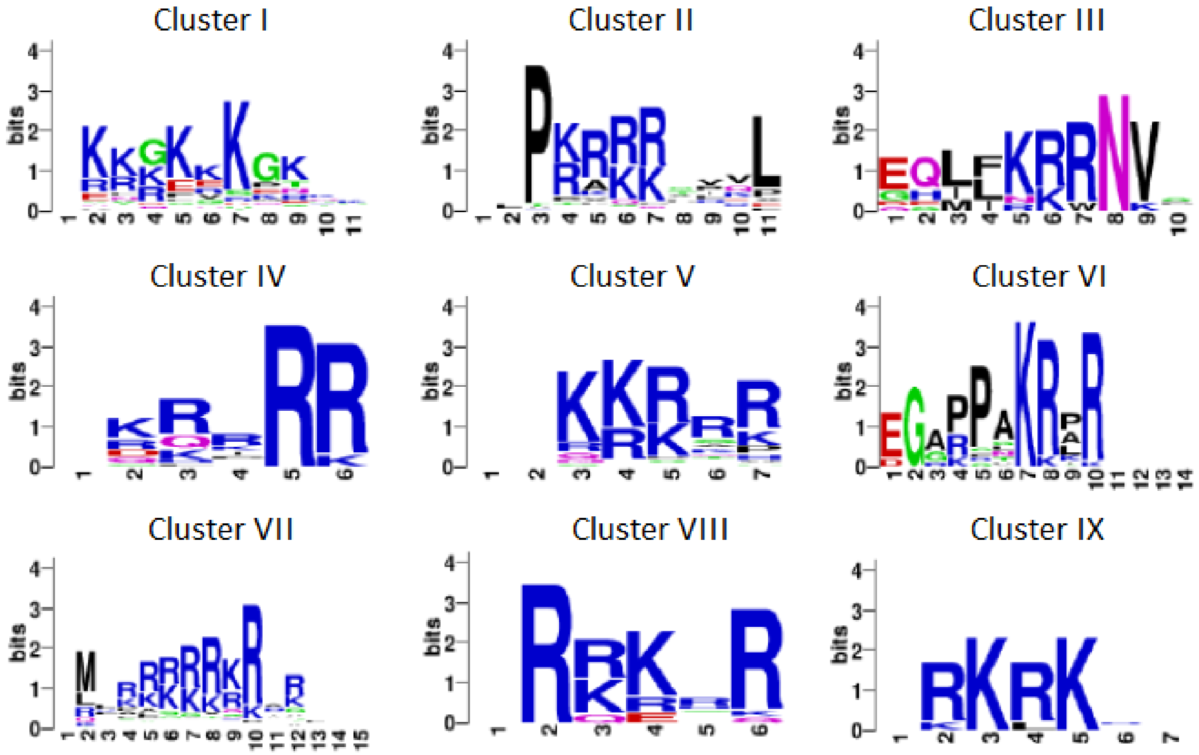


Figure 4: Sequence logo representation of nine randomly chosen monopartite NLS clusters. While sequences of most clusters follow the general "rule" of being a short stretch of basic amino acids for monopartite NLS, there are also exceptions. For example, Clusters II, III and VI contain in addition to highly conserved basic amino acids also highly conserved hydrophobic amino acids. Sequence logos were generated as described in Figure 3.

Sequences of 413 bipartite NLS formed 39 distinct clusters. The phylogenetic tree of these clusters is displayed in Figure S6. Nine randomly chosen clusters of bipartite NLS are shown in Figure 5. Most bipartite signals followed the standard "rule" of two stretches of basic residues separated by a variable linker region. There were, however, also exceptions. For example, Cluster I rather resembles a cluster of monopartite signals: (i) the signal is too short to be bipartite, (ii) it is overall basic and (iii) has no variable linker region. Thus, the signal type annotation of sequences from Cluster I is likely to be wrong. The linker region of Cluster II is dominated by polar and acidic residues. Whereas, the linker region of Clusters III and V has hydrophobic residues conserved. Thus, different patterns of conservation of linker regions might indicate at their different function role, *e.g.* during binding to karyopherins for nuclear import.

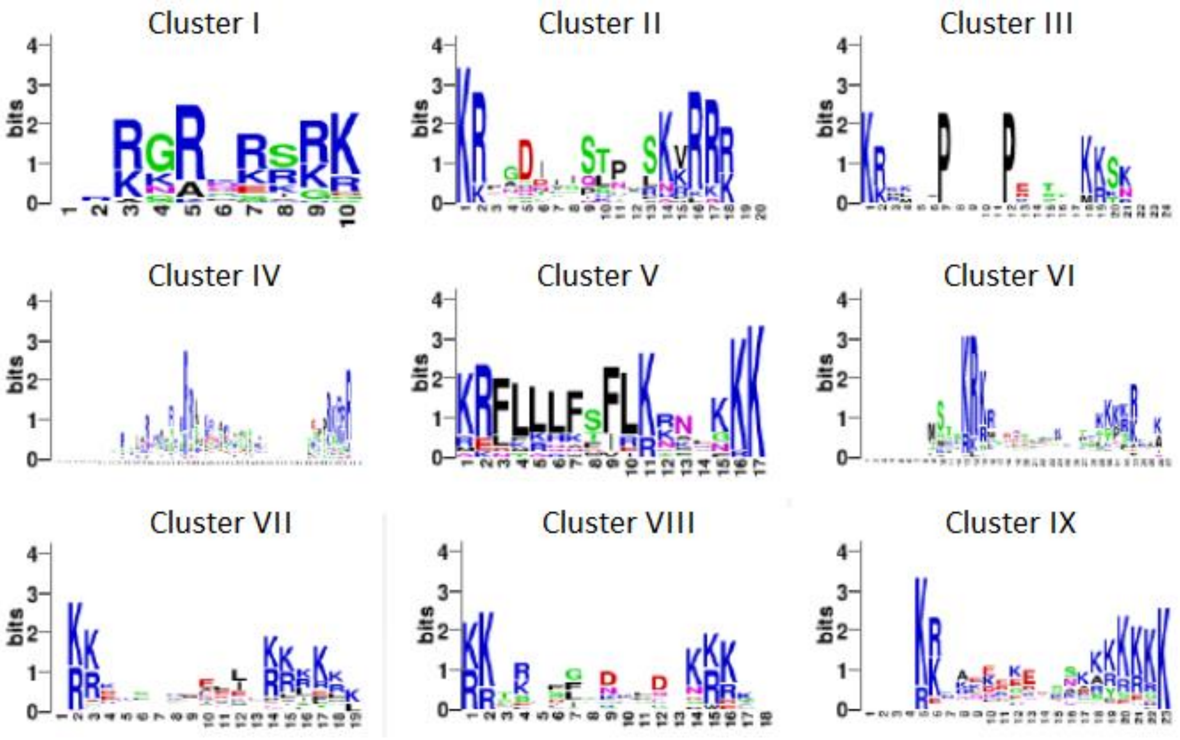


Figure 5: Sequence logo representation of nine randomly chosen bipartite NLS clusters. Similar to monopartite signals, sequences of most bipartite NLS follow the general "rule" of two short stretches of basic amino acids separated by a variable linker region, though there are also exceptions. For example, Cluster I resembles monopartite NLS, and Clusters III and V have conserved hydrophobic residues in their linker region. Sequence logos were generated as in Figure 4.

We had 19 PY-NLS annotated in our data set. Unfortunately, this set was too small to detect reliable consensus sequences for PY-NLS. The phylogenetic tree divided PY-NLS in five clusters and from the sequence logo of its largest cluster (7 sequences), the strongly conserved proline and tyrosine at the C-terminal region could be seen, as well as basic histidine and arginine at the N-terminal region (Figure S7).

Finally, the phylogenetic tree of 817 NES, divided signal sequences in 27 different clusters (Figure S8). The sequence logo of six randomly chosen clusters is presented in Figure 6. NES seemed overall to be less conserved than NES, but overall richer in leucine and other hydrophobic, acidic and polar residues. The, for NLS specific, basic residues were rare in our set of trusted NES.

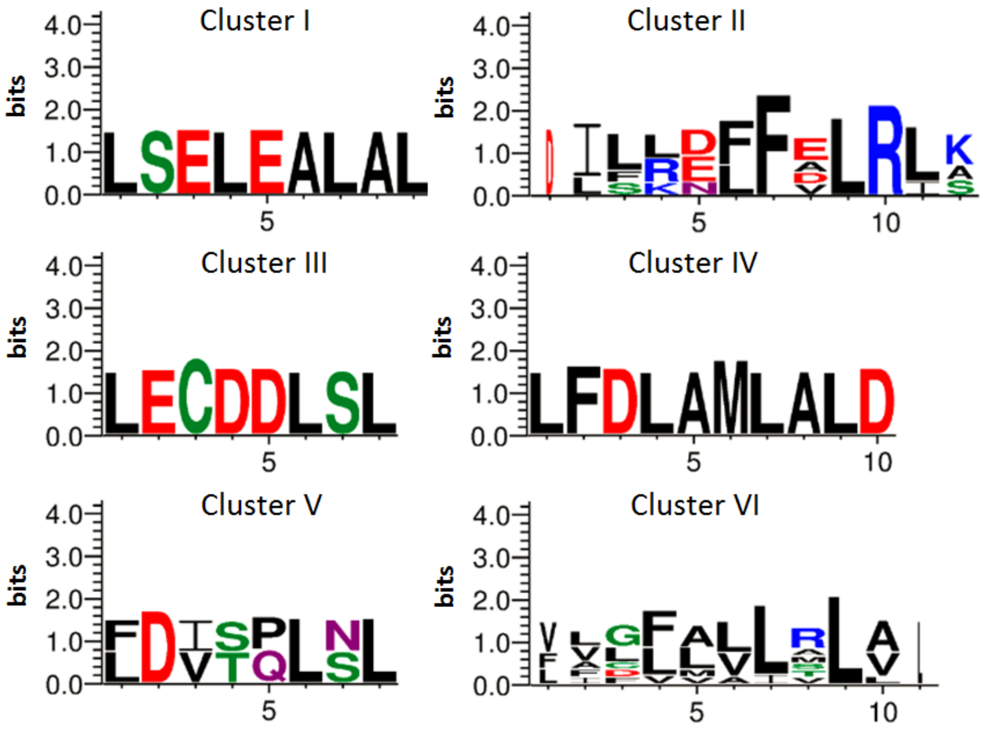


Figure 6: Sequence logo representation of six randomly chosen NES clusters. In contrast to NLS, the sequences of NES are less conserved. In their chemical properties NES are also different from NLS in being mostly built of acidic, hydrophobic and polar residues. Sequence logos were generated as described in Figure 4.

NLSdb 2.0: NES are new and the number of NLS has grown 21-fold

The data set of experimentally determined nuclear localization signals, collected in 2000 for the first version of NLSdb, contained 114 signals. Fifteen years later, the new dataset of trusted, experimentally and by experts annotated, samples contained 2,391 NLS. This number is a 21-fold increase to the data set size from 2000. In addition, the new data set holds sequences of 817 NES, which have not part of the first version of NLSdb.

We applied the “in *silico* mutagenesis” approach [16] to our set of trusted samples. During mutagenesis, we mutated or removed amino acids at different positions of nuclear signals from our trusted set and monitored their matches in nuclear proteins (true signals) and in non-nuclear proteins (false signals). We discarded any potential signal matching in non-nuclear proteins. By doing so, we increased our data set by 4,310 novel potential NLS and by 1,768 novel potential NES.

NLSdb 2.0 vs. NLSdb: increasing coverage from 19 to 50%

We extracted 6,538 sequences of experimentally annotated nuclear proteins from the Swiss-Prot release 2015_01. Of these, 596 proteins (9%) were annotated with NLS. Querying 6,538 proteins with experimental and potential nuclear signals from the first version of NLSdb, we identified NLS in additional 10% of the data (total number of matched proteins was 1,261), thus increasing the coverage from 9% (Swiss-Prot) to 19% (NLSdb). Querying the same data set of nuclear proteins with NLS from NLSdb2.0, we identified signal matches in 3,259 proteins, which correspond to 50% of all proteins. Compared to NLSdb, NLSdb 2.0 increased the coverage in predicting nuclear proteins from 19 to 50%.

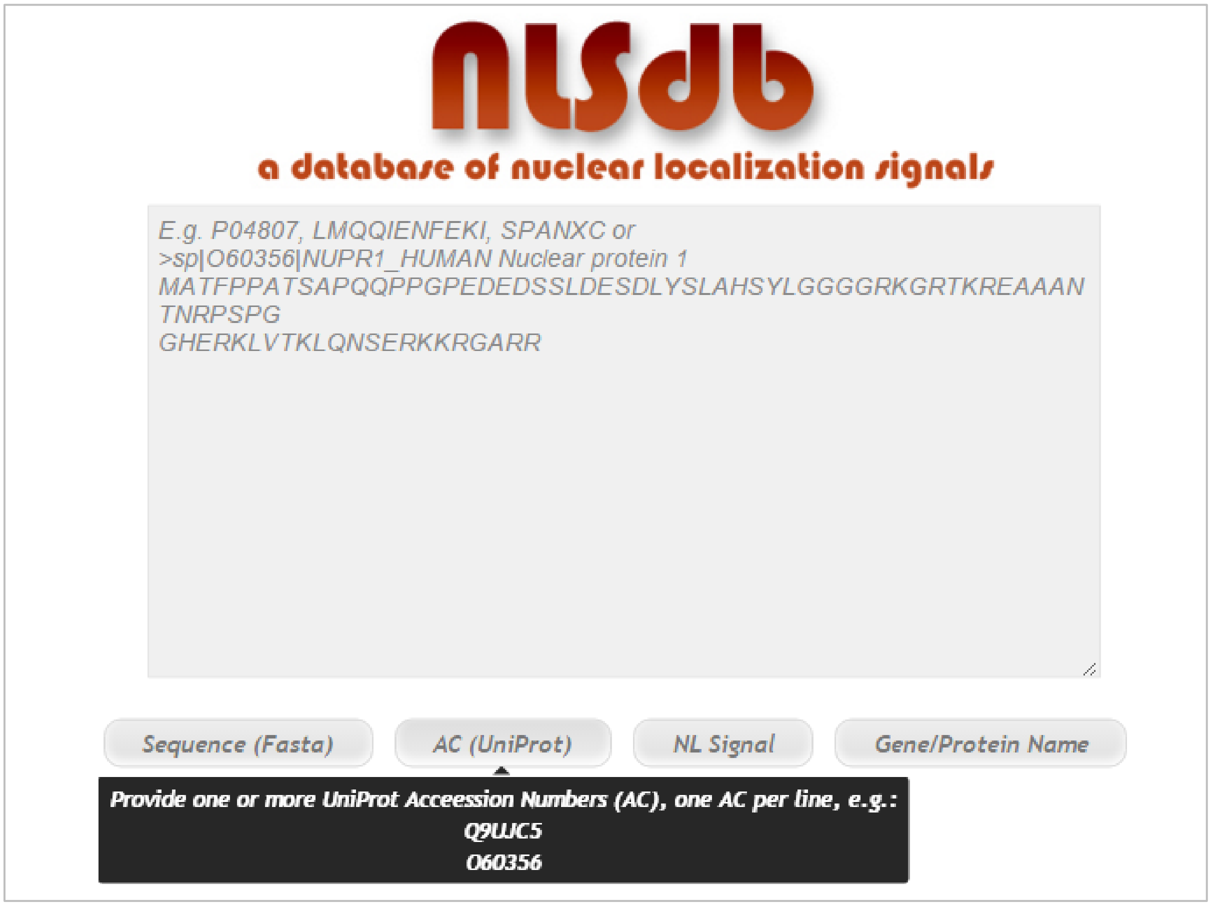
About 5% of nuclear proteins in our set had a Swiss-Prot annotation of NES. Querying nuclear proteins with trusted and potential NES of NLSdb 2.0, we could increase this number by 23%. Thus, the percentage of NES-containing nuclear proteins was 29%.

5.4 Database description

Input formats

The online database of NLSdb 2.0 can be accessed via https://rostlab.org/services/nlsdb2/.

A user can query the database either by *nuclear signals*, to check if his/her signal of interest is contained in our trusted or potential sets, or by *nuclear proteins* to predict the occurrence of NLS and NES in them. Submissions of proteins can be done through providing: (i) their amino acid sequence in FASTA [59] format, (ii) their UniProt [60] accession numbers (ACs), or (iii) their gene and/or protein names (Figure 7).



E.g. P04807, LMQQIENFEKI, SPANXC or
>sp|O60356|NUPR1_HUMAN Nuclear protein 1
MATFPPATSAPQQPPGPEDEDSSLDESDLYSLAHSYLGGGGRKGRTKREAAAN
TNRPSPG
GHERKLVTKLQNSERKKRGARR

Provide one or more UniProt Acceession Numbers (AC), one AC per line, e.g.:
Q9UJC5
O60356

Figure 7: Screenshot of the NLSdb 2.0 submission page. NLSdb 2.0 accepts submissions of four types. Results are returned to the user after clicking one of the four submission buttons which expect: (i) *Sequence (Fasta)*: one or more protein sequences in FASTA [59] format; (ii) *AC (UniProt)*: one or more UniProt [60] accession numbers; (iii) *NL Signal*: one or more sequences of NLS and/or NES; and (iv) *Gene/Protein Name*: names of one or more genes and/or proteins. Hovering with the mouse over the submission buttons displays information about the expected format of a submission (as shown by the black box). Example queries are by default provided also in the text input field.

Data Output

Submissions of nuclear signals (NLS and NES) are simple lookups in our sets of trusted and potential signals. Submissions of protein sequences in FASTA format trigger the matching of all signals stored in our database in sequences of query proteins. Submissions of proteins as UniProt ACs fetch their corresponding FASTA sequences from UniProt to process them as if FASTA sequences were provided as input. Finally, submissions of protein/gene names map these to UniProt ACs, based on pairs of protein/gene names and ACs downloaded from UniProt (version 2012_10), and process them as if UniProt ACs were provided as input. For each query, the output is organized in eleven following fields:

(i) *NL Signal*: amino acid sequence of the query signal.

(ii) *Signal type*: the signal is a monopartite NLS, bipartite NLS, PY-NLS, NES, potential NLS or potential NES.

(iii) *ConfidenceNuc*: the number of structural families of nuclear proteins the signal can be found in.

(iv) *ConfidenceNuc*: the number of structural families of non-nuclear proteins the signal can be found in.

(v) *Annotation Type:* whether the signal annotation is based on experimental findings or it is derived through "*in silico* mutagenesis". For experimentally determined signals, the source of annotation is provided, if available. The source can be the PubMed identifier or the UniProt accession number of the experimental evidence.

(vi) *UniProtKB AC*: UniProt accession number of the source protein(s) the signal is annotated to localize.

(vii) *Start*: start position of the signal in the annotated protein.

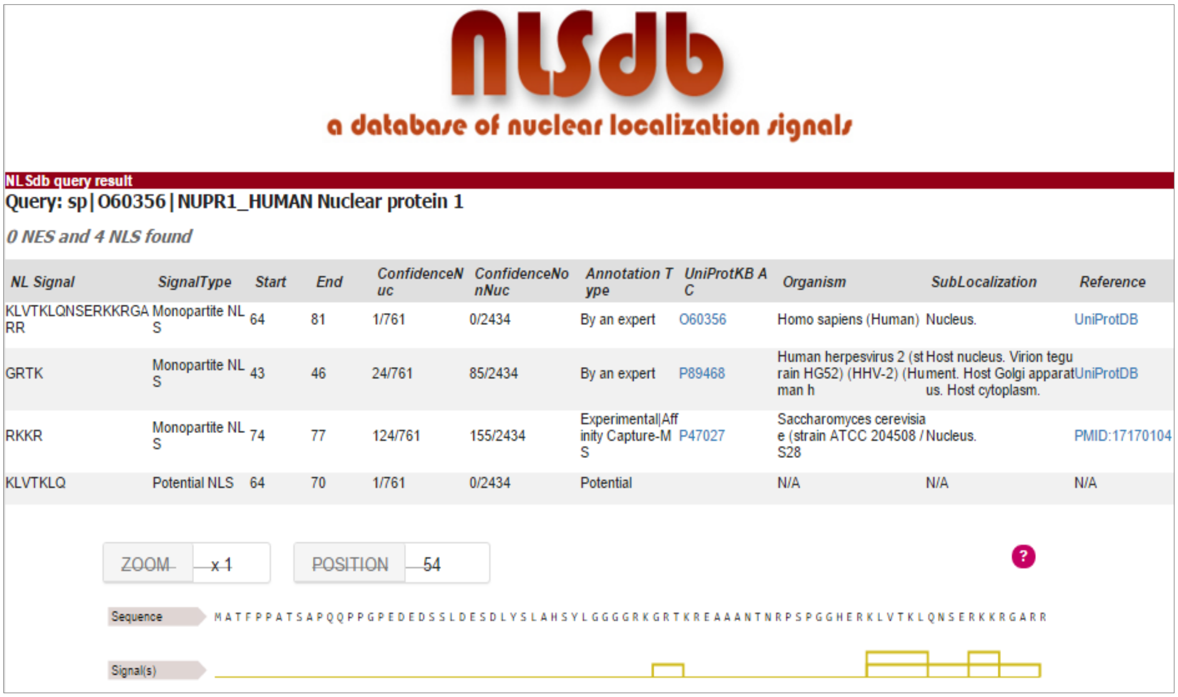
(viii) *End*: end position of the signal in the annotated protein.

(ix) *Organism*: organism annotation of the source protein.

(x) *SubLocalization*: sub-cellular localization annotation of the source protein, extracted from Swiss-Prot, if available.

(xi) *Reference:* The source of the signal annotation, provided for experimentally determined signals only. The source is provided as an active link to either the PubMed article or the database source.

For protein submissions, the annotations of identified nuclear signals are also supported by a graphical visualization. Figure 8 shows NLSdb 2.0 result for the human nuclear protein 1 (UniProt AC: O60356). The protein is identified to contain 0 NES and 4 NLS, of which 1 is a potential signal and 3 are experimentally derived monopartite NLS. These NLS are of human, H*uman herpesvirus 2* and yeast origins. Both yeast and virus NLS are 4 amino acids long and are frequent matches in sequences of other nuclear and non-nuclear proteins. The sequence of the input protein is shown below the results table and the schematic representation of identified NLS is shown below the sequence. Three of four signals (yellow rectangles) match the C-terminal region of the query protein. The longest of three signals is the virus signal, which overlaps with two other experimentally determined eukaryotic monopartite NLS.



a database of nuclear localization signals

NLSdb query result

Query: sp|O60356|NUPR1_HUMAN Nuclear protein 1

0 NES and 4 NLS found

NL Signal	SignalType	Start	End	ConfidenceNuc	ConfidenceNonNuc	Annotation Type	UniProtKB AC	Organism	SubLocalization	Reference
KLVTKLQNSERKKRGARR	Monopartite NLS	64	81	1/761	0/2434	By an expert	O60356	Homo sapiens (Human)	Nucleus.	UniProtDB
GRTK	Monopartite NLS	43	46	24/761	85/2434	By an expert	P89468	Human herpesvirus 2 (strain HG52) (HHV-2) (Human h	Host nucleus. Virion tegument. Host Golgi apparatus. Host cytoplasm.	UniProtDB
RKKR	Monopartite NLS	74	77	124/761	155/2434	Experimental|Affinity Capture-MS	P47027	Saccharomyces cerevisiae (strain ATCC 204508 / S28	Nucleus.	PMID:17170104
KLVTKLQ	Potential NLS	64	70	1/761	0/2434	Potential		N/A	N/A	N/A

ZOOM
x 1
POSITION
54
?
Sequence
MATFPPATSAPQQPPGPEDEDSSLDESDLYSLAHSYLGGGGRKGRTKREAAANTNRPSPGGHERKLVTKLQNSERKKRGARR
Signal(s)

Figure 8: Screenshot of the NLSdb 2.0 results page. Shown is the NLSdb 2.0 result for the query human nuclear protein 1 (UniProt AC: O60356). The header of the result page provides the name of the query protein and the number of nuclear signals identified. Below, the results table provides an overview of signal annotations (*e.g.* signal annotation type, position in the query sequence, source protein and organism). Finally, the positions of identified NLS and NES in the sequence of the query protein are visualized at the bottom of the page. For visualization we used the feature-viewer implementation [61] from the BioJS library [62].

5.5 References

1. Freitas N, Cunha C: **Mechanisms and signals for the nuclear import of proteins.** *Current genomics* 2009, **10:**550-557.
2. Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH: **Classical nuclear localization signals: definition, function, and interaction with importin alpha.** *The Journal of biological chemistry* 2007, **282:**5101-5105.
3. Gorlich D: **Nuclear protein import.** *Current opinion in cell biology* 1997, **9:**412-419.
4. Hodel MR, Corbett AH, Hodel AE: **Dissection of a nuclear localization signal.** *The Journal of biological chemistry* 2001, **276:**1317-1325.
5. Bickmore WA, Sutherland HG: **Addressing protein localization within the nucleus.** *The EMBO journal* 2002, **21:**1248-1254.
6. Lee BJ, Cansizoglu AE, Suel KE, Louis TH, Zhang Z, Chook YM: **Rules for nuclear localization sequence recognition by karyopherin beta 2.** *Cell* 2006, **126:**543-558.
7. Wen W, Meinkoth JL, Tsien RY, Taylor SS: **Identification of a signal for rapid export of proteins from the nucleus.** *Cell* 1995, **82:**463-473.
8. Fischer U, Huber J, Boelens WC, Mattaj IW, Luhrmann R: **The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs.** *Cell* 1995, **82:**475-483.
9. Bogerd HP, Fridell RA, Benson RE, Hua J, Cullen BR: **Protein sequence requirements for function of the human T-cell leukemia virus type 1 Rex nuclear export signal delineated by a novel in vivo randomization-selection assay.** *Molecular and cellular biology* 1996, **16:**4207-4214.
10. la Cour T, Gupta R, Rapacki K, Skriver K, Poulsen FM, Brunak S: **NESbase version 1.0: a database of nuclear export signals.** *Nucleic acids research* 2003, **31:**393-396.
11. la Cour T, Kiemer L, Molgaard A, Gupta R, Skriver K, Brunak S: **Analysis and prediction of leucine-rich nuclear export signals.** *Protein engineering, design & selection : PEDS* 2004, **17:**527-536.
12. Fu SC, Imai K, Horton P: **Prediction of leucine-rich nuclear export signal containing proteins with NESsential.** *Nucleic acids research* 2011, **39:**e111.
13. Stewart M: **Molecular mechanism of the nuclear protein import cycle.** *Nature reviews Molecular cell biology* 2007, **8:**195-208.
14. Cingolani G, Bednenko J, Gillespie MT, Gerace L: **Molecular basis for the recognition of a nonclassical nuclear localization signal by importin beta.** *Molecular cell* 2002, **10:**1345-1353.
15. Nair R, Carter P, Rost B: **NLSdb: database of nuclear localization signals.** *Nucleic acids research* 2003, **31:**397-399.
16. Cokol M, Nair R, Rost B: **Finding nuclear localization signals.** *EMBO reports* 2000, **1:**411-415.
17. Dissanayake K, Toth R, Blakey J, Olsson O, Campbell DG, Prescott AR, MacKintosh C: **ERK/p90(RSK)/14-3-3 signalling has an impact on expression of PEA3 Ets transcription factors via the transcriptional repressor capicua.** *The Biochemical journal* 2011, **433:**515-525.
18. Wilczynska A, Minshall N, Armisen J, Miska EA, Standart N: **Two Piwi proteins, Xiwi and Xili, are expressed in the Xenopus female germline.** *RNA* 2009, **15:**337-345.
19. Miyatake H, Sanjoh A, Unzai S, Matsuda G, Tatsumi Y, Miyamoto Y, Dohmae N, Aida Y: **Crystal structure of human importin-alpha1 (Rch1), revealing a potential autoinhibition mode involving homodimerization.** *PloS one* 2015, **10:**e0115995.

20. Zuchero JB, Belin B, Mullins RD: **Actin binding to WH2 domains regulates nuclear import of the multifunctional actin regulator JMY.** *Molecular biology of the cell* 2012, **23:**853-863.

21. Sengel C, Gavarini S, Sharma N, Ozelius LJ, Bragg DC: **Dimerization of the DYT6 dystonia protein, THAP1, requires residues within the coiled-coil domain.** *Journal of neurochemistry* 2011, **118:**1087-1100.

22. Hedhili S, De Mattei MV, Coudert Y, Bourrie I, Bigot Y, Gantet P: **Three non-autonomous signals collaborate for nuclear targeting of CrMYC2, a Catharanthus roseus bHLH transcription factor.** *BMC research notes* 2010, **3:**301.

23. Lange A, Mills RE, Devine SE, Corbett AH: **A PY-NLS nuclear targeting signal is required for nuclear localization and function of the Saccharomyces cerevisiae mRNA-binding protein Hrp1.** *The Journal of biological chemistry* 2008, **283:**12926-12934.

24. Lin JR, Hu J: **SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring.** *PloS one* 2013, **8:**e76864.

25. Suel KE, Gu H, Chook YM: **Modular organization and combinatorial energetics of proline-tyrosine nuclear localization signals.** *PLoS biology* 2008, **6:**e137.

26. Kosugi S, Yanagawa H, Terauchi R, Tabata S: **NESmapper: accurate prediction of leucine-rich nuclear export signals using activity-based profiles.** *PLoS computational biology* 2014, **10:**e1003841.

27. Garcia-Santisteban I, Banuelos S, Rodriguez JA: **A global survey of CRM1-dependent nuclear export sequences in the human deubiquitinase family.** *The Biochemical journal* 2012, **441:**209-217.

28. Fu SC, Huang HC, Horton P, Juan HF: **ValidNESs: a database of validated leucine-rich nuclear export signals.** *Nucleic acids research* 2013, **41:**D338-343.

29. Xu D, Farmer A, Collett G, Grishin NV, Chook YM: **Sequence and structural analyses of nuclear export signals in the NESdb database.** *Molecular biology of the cell* 2012, **23:**3677-3693.

30. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic acids research* 2000, **28:**45-48.

31. http://www.evidenceontology.org/Welcome.html.

32. Mika S, Rost B: **UniqueProt: Creating representative protein sequence sets.** *Nucleic acids research* 2003, **31:**3789-3791.

33. Rost B: **Twilight zone of protein sequence alignments.** *Protein engineering* 1999, **12:**85-94.

34. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9:**56-68.

35. Thorne JL, Kishino H, Felsenstein J: **An evolutionary model for maximum likelihood alignment of DNA sequences.** *Journal of molecular evolution* 1991, **33:**114-124.

36. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Computer applications in the biosciences : CABIOS* 1992, **8:**275-282.

37. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.

38. Michener CD, Sokal RR: **A quantitative approach to a problem of classification.** *Evolution* 1957, **11:**490–499.

39. Krause A, Stoye J, Vingron M: **Large scale hierarchical clustering of protein sequences.** *BMC bioinformatics* 2005, **6:**15.

40. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Molecular biology and evolution* 2013, **30:**772-780.

41. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic acids research* 1990, **18:**6097-6100.

42. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome research* 2004, **14:**1188-1190.

43. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22:**1658-1659.

44. D'Amelio E, Salemi S, D'Amelio R: **Anti-infectious human vaccination in historical perspective.** *International reviews of immunology* 2015**:**1-32.

45. Plotkin S: **History of vaccination.** *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111:**12283-12287.

46. Rice SA, Davido DJ: **HSV-1 ICP22: hijacking host nuclear functions to enhance viral infection.** *Future microbiology* 2013, **8:**311-321.

47. Dechtawewat T, Songprakhon P, Limjindaporn T, Puttikhunt C, Kasinrerk W, Saitornuang S, Yenchitsomanus PT, Noisakran S: **Role of human heterogeneous nuclear ribonucleoprotein C1/C2 in dengue virus replication.** *Virology journal* 2015, **12:**14.

48. Lloyd RE: **Nuclear proteins hijacked by mammalian cytoplasmic plus strand RNA viruses.** *Virology* 2015, **479-480:**457-474.

49. Konig R, Stertz S, Zhou Y, Inoue A, Hoffmann HH, Bhattacharyya S, Alamares JG, Tscherne DM, Ortigoza MB, Liang Y, et al: **Human host factors required for influenza virus replication.** *Nature* 2010, **463:**813-817.

50. Bonamassa B, Ciccarese F, Antonio VD, Contarini A, Palu G, Alvisi G: **Hepatitis C virus and host cell nuclear transport machinery: a clandestine affair.** *Frontiers in microbiology* 2015, **6:**619.

51. Le Sage V, Mouland AJ, Valiente-Echeverria F: **Roles of HIV-1 capsid in viral replication and immune evasion.** *Virus research* 2014, **193:**116-129.

52. Radigan KA, Misharin AV, Chi M, Budinger GS: **Modeling human influenza infection in the laboratory.** *Infection and drug resistance* 2015, **8:**311-320.

53. http://www.who.int/mediacentre/factsheets/fs211/en/.

54. Trooskin SB, Reynolds H, Kostman JR: **Access to Costly New Hepatitis C Drugs: Medicine, Money, and Advocacy.** *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2015, **61:**1825-1830.

55. http://www.avert.org/professionals/hiv-science/types-strains.

56. Manzardo C, Guardo AC, Letang E, Plana M, Gatell JM, Miro JM: **Opportunistic infections and immune reconstitution inflammatory syndrome in HIV-1-infected adults in the combined antiretroviral therapy era: a comprehensive review.** *Expert review of anti-infective therapy* 2015, **13:**751-767.

57. Smith CJ, Ryom L, Weber R, Morlat P, Pradier C, Reiss P, Kowalska JD, de Wit S, Law M, el Sadr W, et al: **Trends in underlying causes of death in people with HIV from 1999 to 2011 (D:A:D): a multicohort collaboration.** *Lancet* 2014, **384:**241-248.

58. Stewart A, Chan Carusone S, To K, Schaefer-McDaniel N, Halman M, Grimes R: **Causes of Death in HIV Patients and the Evolution of an AIDS Hospice: 1988-2008.** *AIDS research and treatment* 2012, **2012:**390406.

59. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proceedings of the National Academy of Sciences of the United States of America* 1988, **85:**2444-2448.

60. **UniProt: a hub for protein information.** *Nucleic acids research* 2015, **43:**D204-212.

61. https://github.com/calipho-sib/feature-viewer.

62. Corpas M, Jimenez R, Carbon SJ, Garcia A, Garcia L, Goldberg T, Gomez J, Kalderimis A, Lewis SE, Mulvany I, et al: **BioJS: an open source standard for biological visualisation - its status in 2014.** *F1000Research* 2014, **3:**55.

6 pEffect: prediction of bacterial type III effector proteins

6.1 Introduction

The type III secretion system is a key mechanism for the transport of effector proteins of pathogenic and endosymbiotic Gram-negative bacteria into the cytoplasm of host cells [1-5]. During infection, effectors convert host resources to work to bacterial advantage. Previous computational methods for the prediction of type III effectors have mainly employed information encoded in the N-terminal sequence [6-9], as it contains most important signals that govern the translocation of effectors through the type III secretion machinery [1]. An independent, recent benchmark study showed that current state-of-art-methods predict type III effectors at comparable levels of at best 80% accuracy and 80% coverage [10]; thus, there still seems to be room for substantial improvement.

In this Chapter, a new method, pEffect, is introduced that predicts type III effector proteins from the information encoded in the entire protein amino acid sequence. It combines sequence similarity-based inferences (PSI-BLAST [11]) with *de novo* predictions using machine learning (Profile Kernel Support Vector Machines [12-14]). To allow users to focus on most relevant results, pEffect provides a score reflecting the strength of each prediction. The method was developed using a positive data set comprising type III effectors extracted from literature and UniProt [15] and a negative data set combining bacterial non-effector proteins and effector sequence-similar eukaryotic proteins. Tested on a non-redundant test set, pEffect reaches high levels of 87±7% accuracy and 95±5% coverage. The method importantly improves over its competitors, boosting performance by at least 7% for bacterial effectors and as much as 3-fold on data sets containing eukaryotic proteins. This result suggests that the information required for distinguishing effectors is not confined to any particular part of the amino acid sequence, but is rather distributed over the entire protein sequence. This biological feature helps pEffect to maintain a high level of accuracy even when tested on sequence fragments. pEffect can thus be effectively applied directly to metagenomic read data, facilitating studies of microbial community interactions. Applied to proteomes of all fully sequences prokaryotic organisms, pEffect identifies a wide variety of recently evolved effectors. These highlight the possibility of a type III secretion ancestor dating to times prior to the archaea/bacteria split. pEffect is available as a public web server and as a standalone version for download at http://www.bromberglab.org/services/pEffect.

6.2 Materials and Methods

Data sets for development and evaluation

Our positive data set of known type III effector proteins was extracted from literature [6, 16-23] and the Pseudomonas-Plant Interaction web site [24]. The corresponding amino acid sequences were taken from the UniProt database [15], 2012_01 release. We additionally queried UniProt with keywords ‘type III effector’, ‘type three effector’ and ‘T3SS effector’ and manually curated the results for experimentally identified effectors. Our positive data set comprised 1,388 proteins.

To compile our negative data set of non-type III effectors we used experimentally annotated Swiss-Prot proteins [25], 2012_01 release. We extracted all bacterial proteins that were NOT annotated as type III effectors and had no significant sequence similarity (BLAST [26] E-value > 10) to any type III effector in our positive set. We also added all eukaryotic proteins applying no sequence similarity filters. Our negative set contained roughly 470,000 proteins.

We removed from our sets all proteins annotated as ‘uncharacterized’, ‘putative’, or ‘fragment’. We reduced sequence redundancy independently in each set using UniqueProt [27], ascertaining that no pair of proteins in one set had alignment length of less than 35 residues or a positive HSSP-value (HVAL ≥ 0) [28, 29]. After redundancy reduction our sequence-unique sets contained 115 type III effector proteins from 43 different bacterial species and 3,460 non-effector proteins (of which 37% were bacterial). Here, we term this set of sequences (positive and negative sets together) as the *Development set*. All pEffect performance results reported here across the Development set and its subsets are based on five-fold cross-validation experiments, *i.e.* we split the entire set into five similarly-sized subsets and trained five models, each on a different combination of four of these subsets, and tested each model on every subset exactly once.

Data sets for additional testing

We benchmarked pEffect against other methods using the following data sets:

(1) We collected all type III effectors added to UniProt between releases 2012_01 and 2014_08 and non-type III bacterial and eukaryotic proteins added between same releases to

Swiss-Prot. These were redundancy reduced at HVAL< 0 to produce the *UniProt'14*$_{HVAL0}$ test set (107 effectors and 1,159 non-effectors).

(2) To answer the question "how well will pEffect perform on protein sequences added to databases within the next six months?" we collected the proteins added to UniProt (type III effectors) and Swiss-Prot (non-effector bacterial and eukaryotic sequences) after the 2014_08 release, producing the set *UniProt'15*$_{Full}$ (498 effectors and 1,509 non-effectors).

(3) We also extracted all bacterial type III effectors from the T3DB database [30] – *T3DB*$_{Full}$ set (218 effectors and 831 non-effectors). We deliberately kept the redundancy in this set (up to HVAL = 66, *i.e.* over 85% pairwise sequence identity over 450 residues aligned).

(4) Finally, we redundancy reduced T3DB set at HVAL<0. This gave the *T3DB*$_{HVAL0}$ set (66 effectors and 128 non-effectors).

Prediction method pEffect

Inspired by the high prediction performance of LocTree3 [31] (Chapter 3), pEffect similarly combined homology-based predictions if available and *de novo* predictions otherwise:

Sequence similarity-based predictions: We transferred type III effector annotations by homology using PSI-BLAST [11] alignments. For every query sequence we generated a PSI-BLAST profile (two iterations, inclusion threshold E-value ≤ 10-3) using an 80% non-redundant database combining UniProt [15] and PDB [32]. We then aligned this profile (inclusion E-value ≤ 10-3) against all type III effectors in our Development set (1,388 proteins). For known effectors, we excluded PSI-BLAST self-hits. We transferred annotation to the query protein from the hit with the highest pairwise sequence identity of all retrieved alignments.

De novo predictions: We used the Support Vector Machine (SVM) [12] implementation of WEKA [33] and the Profile Kernel function [13, 14] (Chapter 2) to discriminate between type III effector and non-effector proteins. We found the Profile Kernel parameters k=4 and σ=7 to provide best results. Note we determined the parameters for the SVM and the Profile Kernel separately for each fold in our 5-fold cross-validation and, thus, never optimized them on the test sets.

State-of-the-art predictors for type III effector proteins

We used the following state-of-the-art methods with their default parameters that predict bacterial type III effector proteins and that are publicly accessible:

1. BPBAac [7] uses an SVM to predict type III effectors. Predictions are based on the position-specific amino acid composition (Aac) profiles within 100 N-terminal residues of a protein sequence. BPBAac was trained on non-redundant sets of 154 type III effectors curated manually from literature and 308 non-effectors randomly selected from various bacteria, followed by removal of the known effectors and their homologs. BPBAac is available at http://biocomputer.bio.cuhk.edu.hk/softwares/BPBAac.
2. Effective T3 [6] applies the Naïve Bayes classification to predict type III effectors on the basis of various features of the 25 N-terminal residues, including frequencies of amino acids, short peptides, and residues with certain physico-chemical properties. Effective T3 was trained on a positive set of 100 manually curated type III effectors from literature. The negative set of 200 non-effector proteins was collected by randomly choosing proteins from animal and plant pathogens, omitting known effectors. Effective T3 is available at http://www.effectors.org/.
3. T3_MM [9] is based on BPBAac and uses Aac profiles of adjacent residues to predict type III effectors. It employs a Markov model to calculate the Aac probability difference between type III effector and non-effector proteins. T3_MM was trained on BPBAac training data. Predictions are made using 100 N-terminal residues. T3_MM is available at http://biocomputer.bio.cuhk.edu.hk/T3DB/T3_MM.php.

T3DB ortholog clusters of the type III secretion system (T3SS) machinery

T3DB is a database of experimentally annotated T3SS-related proteins in 36 bacterial taxa. Proteins of the same function and the same evolutionary origin are clustered in T3DB into *T3 Ortholog clusters* [34]. The proteins of these clusters form ten components of the T3SS. Proteins of five of these components (export apparatus, inner membrane ring, outer membrane ring, cytoplasmic ring, and ATPase) are present in all 36 taxa in T3DB. We thus defined the minimum number of five components necessary for the formation of the T3SS machinery. Four of these, with the exception of the outer membrane ring, have also been defined as core in [35].

Evolutionary distances

We extracted evolutionary distances from the phylogenetic tree in the Newick format of 2,966 bacterial and archaeal taxa, which has been inferred from 38 concatenated genes [36].

6.3 Results

pEffect: high cross-validated performance of F1 = 0.91

The accuracy of the PSI-BLAST sequence similarity-based inference, *i.e.* a look up for a sequence-similar experimentally annotated type III effector protein, was comparable to that of our *de novo* prediction method on the cross-validated Development set (Table 1: 91% vs. 92%). However, its coverage was significantly higher (84% vs. 60%). This result encouraged us to use a simple protocol, introduced in our recent work, LocTree3 [31], that unites PSI-BLAST whenever possible (Table 1: F1 = 0.87 on the complete Development set) and the SVM if no PSI-BLAST results were available (Table 1: F1 = 0.67 on proteins with no PSI-BLAST hit). The combined method, pEffect, outperformed both its components, reaching an F1 measure of 0.91 (Table 1).

Method	*True Positives*	*False Negatives*	*False Positives*	*True Negatives*	*Acc*[5]	*Cov*[5]	*F1*[5]
PSI-BLAST[1]	97	18	10	3450	91±7	84±8	0.87±0.09
De novo[2]	69	46	6	3454	92±8	60±11	0.73±0.11
De novo$_{\textit{No\_PSI-BLAST\_hit}}$[3]	12	6	6	3444	67±25	67±28	0.67±0.23
pEffect[4]	109	6	16	3444	87±7	95±5	0.91±0.08

Table 1: Performance of pEffect and its components on the Development set

[1] PSI-BLAST: sequence similarity-based inference component of pEffect on all 3,755 proteins of the full Development set.

[2] *De novo*: SVM-based prediction component on the full Development set.

[3] *De novo*$_{\text{No\_PSI-BLAST\_hit}}$: SVM-based prediction component tested on the set of 3,468 proteins that did not align to any effector using PSI-BLAST.

[4] *pEffect*: PSI-BLAST predictions, if available, and *de novo* otherwise on the full Development set.

[5] Performance measures: Acc, accuracy; Cov, coverage; '±' standard errors (Chapter 2.1); F1=2·Acc·Cov/(100·[Acc+Cov]). Highest value in each column is in bold.

pEffect outperforms other methods

We compared pEffect's performance to the publicly available methods: BPBAac [7], Effective T3 [6] and T3_MM [9]. In contrast to pEffect, all these methods focus exclusively on the protein's N-terminal sequence features. BPBAac and T3_MM rely solely on amino acid composition, while Effective T3 combines amino acid composition and secondary structure information. We compared the prediction performance of these methods to pEffect on UniProt protein sequences, which were NOT used for the development of any method, and on T3DB proteins, some of which were used for the development of all methods, including pEffect. Our method outperformed its competitors on all data sets (Figure 1A). Interestingly, the F1 performance of pEffect was at least 0.58 higher than of other methods when tested on any data set containing eukaryotic proteins (0.58 difference T3_MM vs. pEffect on both UniProt sets). Thus, pEffect is the most accurate method in distinguishing type III effectors from other bacterial sequences (F1 > 0.64) and from eukaryotic sequences (F1 > 0.85). The latter ability will be important when considering, for example, sequences from unfiltered metagenomic samples [37].

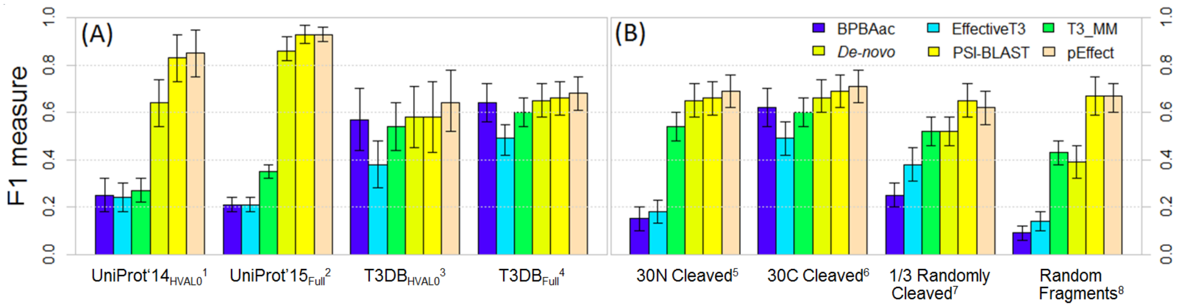


Figure 1: pEffect benchmarking against other methods. We measured the performance of BPBAac [7], EffectiveT3 [6], T3_MM [9] and our own method, pEffect, using the F1 measure (Table 1). We also measured F1 for *de novo* (SVM-based) and PSI-BLAST predictions alone. **Panel (A)** shows performance on additional data sets for testing, which include:

[1] UniProt'14$_{HVAL0}$: 107 effectors and 1,159 non-effector bacterial and eukaryotic proteins, added to UniProt between releases 2012_01 and 2014_08, sequence homology reduced at HVAL< 0

[2] UniProt'15$_{Full}$: 498 effectors and 1,509 non-effector bacterial and eukaryotic proteins added to UniProt after 2012_08 release, NOT homology reduced

[3] T3DB$_{HVAL0}$: 66 effectors and 128 non-effector bacterial proteins from the T3DB database, sequence homology reduced at HVAL < 0

[4] T3DB$_{Full}$: 218 effectors and 831 non-effector bacterial proteins from T3DB, NOT homology reduced

Panel (B) shows performance on protein fragments produced from the T3DB$_{Full}$[4] set, which include:

[5] 30N Cleaved: 30 N-terminal amino acids cleaved off

[6] 30C Cleaved: 30 C-terminal amino acids cleaved off

[7] 1/3 Randomly Cleaved: randomly selected one third of amino acids cleaved off

[8] Random Fragments: randomly selected fragments of a typical translated read length (Figure 2)

pEffect maintains high performance even for sequence fragments

To evaluate pEffect's ability to annotate effectors from incomplete sequences, we fragmented the proteins from the T3DB$_{Full}$ set – the set for which methods developed by others performed best (Figure 1A). We used four different approaches to generate protein fragments: (i) retaining the entire protein sequence, but removing 30 N-terminal residues, (ii) retaining the entire protein sequence, but removing 30 C-terminal residues, (iii) randomly removing one third of residues for each protein sequence and (iv) randomly picking from each sequence a single fragment of a typical translated read length (Figure 2).

pEffect outperformed all external methods for all types of protein fragments (Figure 1B). All methods, as expected from their training, performed best on the C-term cleaved fragments (approach ii). The worst performance was for random sequence fragments (approach iv). Interestingly, the performance for pEffect changed insignificantly from F1 = 0.69 to F1 = 0.67 on the random fragments set. In general, for all fragment sets the pEffect and PSI-BLAST performances were within the standard error of what was obtained using full-length sequences (T3DB$_{Full}$ set; Figure 1A). These results suggest that the features distinguishing type III effectors are spread over the entire protein sequence and are picked up PSI-BLAST or the more advanced *k*-mer comparisons of the SVM Profile Kernel.

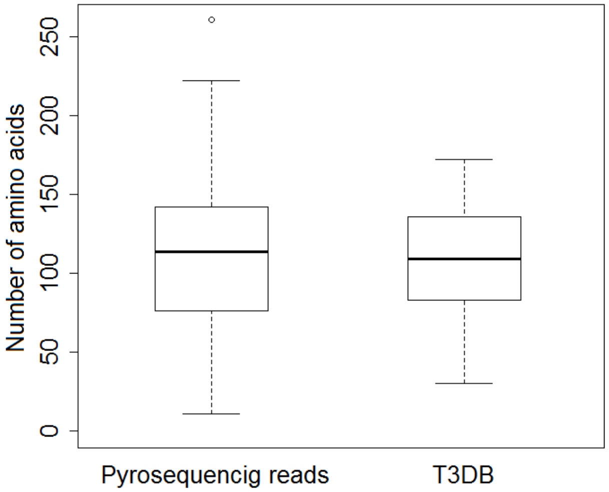


Figure 2: Distribution of a typical translated read length. "Pyrosequencing reads": amino acid lengths of open reading frames translated (between start and stop codons) from eight different snow and soil-collected metagenomic data sets (collaborator data) using the getorf [38] program. "T3DB": amino acid lengths of randomly picked fragments (one fragment per sequence) from the T3DB$_{Full}$ set. The distribution of translated read lengths in the T3DB set follows the distribution of read lengths in "real" metagenomic samples and averages at 110 amino acids.

Reliability index provides more confidence in predictions

pEffect provides a reliability index (RI) to measure the confidence of a prediction. RI is a value between 0 and 100, with 100 denoting most confident predictions. For PSI-BLAST searches, RIs are normalized values of percentage pairwise sequence identities read of the alignments. For *de novo* predictions, RIs are values corresponding to SVM scores. Sampling at lower RIs results in a higher number of predicted samples, though at reduced accuracy. Higher accuracy predictions are obtained by sampling at higher RIs, thus reducing the total number of predicted samples. For example, at the default threshold of RI > 50, over 87% of all predictions of type III effectors are correct and of all effectors in our set 95% are identified (Figure 3: black arrow). At a higher reliability index, RI > 80, effector predictions are correct 96% of the time, but only 78% of all effectors in the set are identified (Figure 3: gray arrow). Thus, a user can make a choice for the reliability of a prediction that is most fitting to his or her purposes: identifying more effectors at lower accuracy or fewer high confidence effectors. Moreover, he or she can focus only on *de novo* predictions (*i.e.* of new, previously unseen, effectors) or on PSI-BLAST predictions (*i.e.* validated homologs of known effectors), as the source of a prediction is provided for each result.

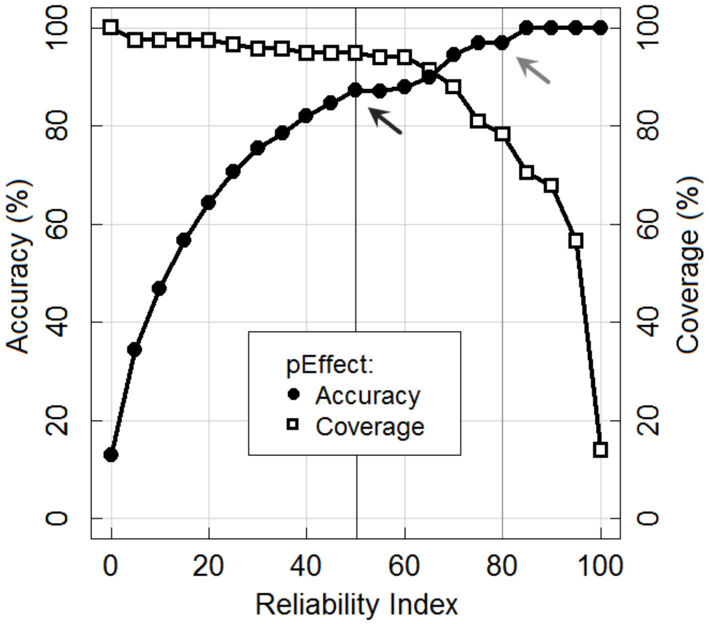


Figure 3: Reliable predictions are more accurate. The figure shows the cumulative percentage of Accuracy/Coverage (Chapter 2) of pEffect's predictions at or above a given reliability index (RI). The graphs were obtained using the Development set of 115 type III effector and 3,460 non-effector proteins in a five-fold cross-validation. At the default reliability score of RI = 50 (black vertical line), 95% of type III effectors are identified at 87% accuracy (black arrow). At a higher RI = 80 (gray vertical line), prediction accuracy increases to 97% at the cost of lower coverage of 78% (gray arrow).

Type III effectors prediction in full proteomes

We used pEffect to annotate type III effectors in the proteomes of fully sequenced 862 bacterial (274 gram-positive and 588 gram-negative bacteria) and 90 archaeal organisms downloaded from the European Bioinformatics Institute (EBI, [39]).

pEffect predicted each bacterium to contain at least one type III effector (Figure 4; a minimum of 0.8% of a proteome is predicted as effectors). For some gram-negative bacteria over 750 type III effectors were predicted (*e.g.* 1,207 effectors in *Sorangium cellulosum* So ce56, 870 effectors in *Stigmatella aurantiaca* DW4/3-1, 826 effectors in *Corallococcus coralloides* DSM 2259 and 792 effectors in *Haliangium ochraceum* DSM 14365). *Stigmatella aurantiaca* DW4/3-1 is hypothesized to have the type III secretion system and its effectors [40]. For the other three species we could not find any literature record.

Overall, the number of predicted type III effectors ranged between 1% and 15% of the whole proteome in gram-negative bacteria, and between 1% and 10% in gram-positive bacteria (Figure 4). To further understand our predictions, we retrieved UniProt keywords of predicted effectors. Their annotations varied widely (Table 2), with the most common for both types of bacteria being transferase, depicting a large class of enzymes that are responsible for the transfer of specific functional groups from one molecule to another, nucleotide-binding, a common functionality of effector proteins, ATP-binding that is also an essential component of the type III secretion system (T3SS), and kinase, which is necessary for the expression of the T3SS genes. About one fourth (26-29% per proteomes) of predicted type III effectors were functionally "unknown" (Table 2).

Interestingly, we also predicted type III effectors in all archaeal proteomes, with over 100 effectors identified in the proteomes of *Haloterrigena turkmenica* DSM 5511 and *Methanosarcina acetivorans* C2A (126 and 105 effectors, respectively). On average, there were fewer effectors predicted in archaea than in bacteria: 1.9% is the overall per-organism number for archaea vs. 3.4% for gram-positive and 4.6% gram-negative bacteria (Figure 4). The most frequent annotations of predicted archaeal effectors were similar to those for predicted bacterial effectors, namely "unknown", nucleotide-binding, ATP-binding and transferase (Table 2).

	UniProt keywords (PSI-BLAST predictions)	*Frequency*	*UniProt keywords (SVM de novo predictions)*	*Frequency*
ARCHAEA	Uncharacterized protein	29.9%	Uncharacterized protein	40.1%
	Hydrolase	5.6%	Oxidoreducatase	5.1%
	Cytoplasm	5.2%	Plasmid	4.8%
	Nucleotide-binding	5.2%	Transferase	4.1%
	ATP-binding	4.9%	Metal-binding	3.7%
	Metal-binding	4.5%	Flavoprotein	3.6%
	Zinc	4.0%	FAD	3.3%
	Chaperone	4.0%	Lyase	2.4%
BACTERIA (+)	Uncharacterized protein	25.6%	Uncharacterized protein	25.6%
	Transferase	6.4%	Transferase	6.7%
	Hydrolase	6.0%	Nucleotide-binding	6.6%
	Nucleotide-binding	5.3%	ATP-binding	6.5%
	ATP-binding	4.7%	Kinase	3.7%
	Kinase	4.7%	Oxidoreductase	3.7%
	Cytoplasm	4.1%	Phosphoprotein	3.0%
	Serine/threonine-protein kinase	2.7%	Metal-binding	2.3%
BACTERIA (-)	Uncharacterized protein	27.8%	Uncharacterized protein	29.1%
	Hydrolase	4.9%	Transferase	7.6%
	Cytoplasm	4.5%	Nucleotide-binding	5.4%
	Transferase	4.4%	Kinase	5.3%
	Metal-binding	3.9%	ATP-binding	5.3%
	Nucleotide-binding	3.9%	Phosphoprotein	4.7%
	ATP-binding	3.3%	Oxidoreductase	2.4%
	Kinase	3.2%	Membrane	2.1%

Table 2: Top eight most frequent UniProt keywords associated with pEffect's predicted effectors. The table lists top eight most frequent keywords retrieved from UniProt for the proteins predicted as type III effectors in the proteomes of 90 archaeal, 274 gram-positive bacterial and 588 gram-negative bacterial species.

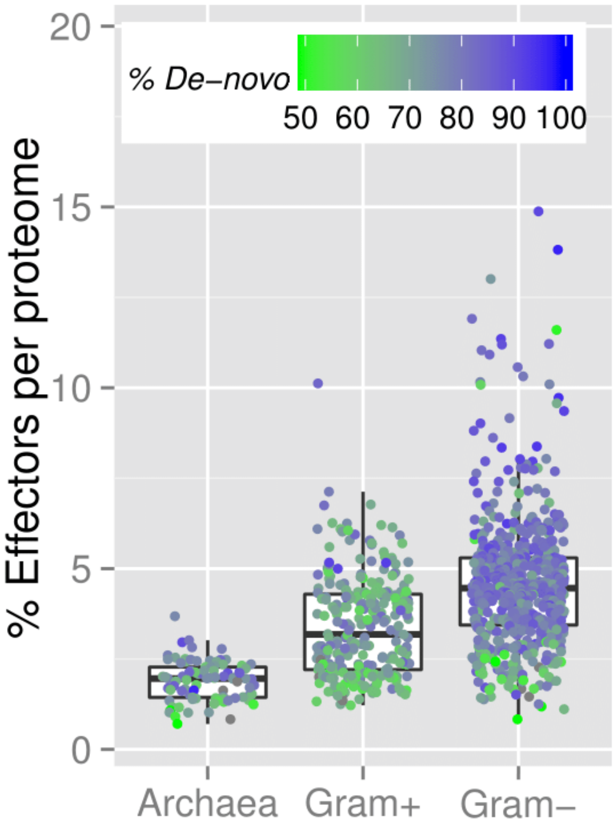


Figure 4: Percentage of predicted effectors in full proteomes. The figure shows the box-plot-and-instance representation of percentages of pEffect's predicted type III effectors (Y-axis) in 90 archaeal, 274 gram-positive and 588 gram-negative bacterial organisms (X-axis), which are shown as dots. At least 50% of effector predictions in all, except 11 organisms in our set were predicted *de novo*. In the figure, the colors represent the percentage of *de novo* predictions for each organism: from green (50% *de novo*, 50% PSI-BLAST) to blue (100% *de novo*, 0% PSI-BLAST). While effectors predicted in archaea and gram-positive bacteria are often picked up by PSI-BLAST, effectors in gram-negative bacteria are mostly *de novo* predictions.

T3SS likely defined by 5 type III machinery components and ≥5% predicted effectors

We aimed to identify those proteomes that are likely to have the type III secretion system (T3SS) machinery. For this, we BLASTed (E-value ≤ 10^{-3}) proteins of five T3DB Ortholog clusters against the full proteomes of our 862 bacteria and 90 archaea set. We found that, as expected, archaea never contain a full T3SS (maximum three out of five components; Figure 5A). In gram-negative bacteria, the number of predicted effectors correlated much better with the number of type III machinery components than in gram-positive bacteria (Figures 5B-C; Pearson's correlation *r*=0.37 and *r*=0.13, respectively). Based on our observations in archaea and gram-positive bacteria, we suggest, as a rule of

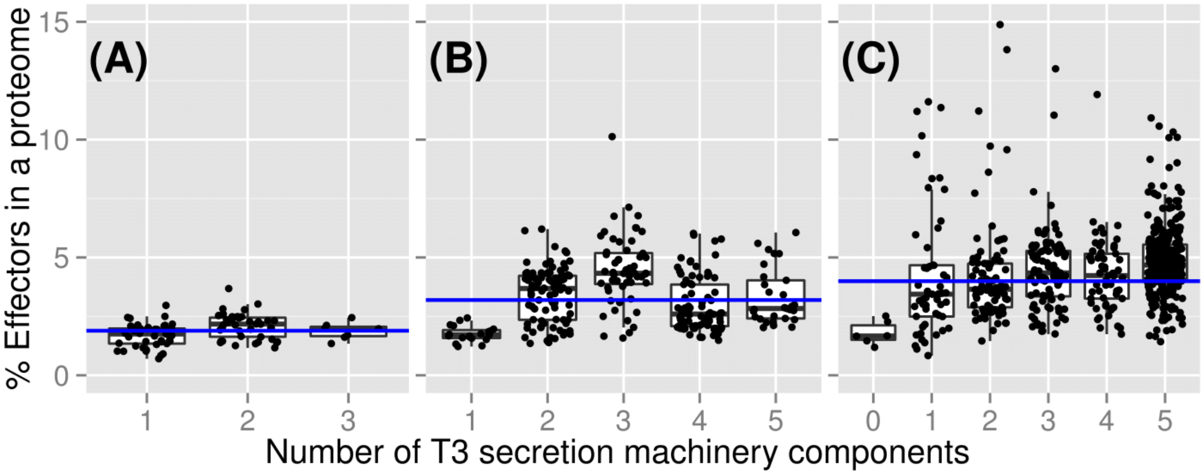


Figure 5: Conservation of T3SS machinery components in full proteomes. (A) 90 archaeal, **(B)** 274 gram-positive and **(C)** 588 gram-negative bacterial proteomes, shown as dots in the figure, were scanned for the presence of T3SS. The percentage of type III effectors per proteome predicted by pEffect (Y-axis) is compared to the number of T3SS machinery components identified in these proteomes (X-axis). While type III effectors compose up to 3.7% of an archaeal proteome (mean 1.9%, blue horizontal line), this number is much larger for bacteria, reaching up to 10% of an entire proteome for gram-positive (mean 3.4%), and 15% for gram-negative bacteria (mean 4.6%). Six gram-negative bacterial species did not contain detectable homologs of any of five machinery components, indicating that their genomes are further diverged than those of other species.

thumb, that the conservation of five type III machinery components and ≥5% of the genome dedicated to effectors provide a strong evidence for the presence of a T3SS in an organism. With these cutoffs, we identified 20% (120 species) of the gram-negative bacteria in our set as type III secreting. No archaeal species and only five gram-positive bacteria fit these cutoffs. We searched the literature for annotation of ten randomly chosen gram-negative bacteria likely to have the T3SS. We found evidence of type III machinery in seven of the ten organisms [41-47]. For three bacteria the secretion machinery has not been studied.

6.4 Discussion

pEffect successfully combines PSI-BLAST and de novo predictions

PSI-BLAST is a commonly used tool for protein function annotation through sequence similarity. PSI-BLAST was first published nearly two decades ago and is continuously improving through growing databases and better alignment techniques [48]. Applied to our sequence-unique Development set, PSI-BLAST correctly annotated type III effector proteins at 91±7% accuracy and 84±8% coverage (F1 = 0.87 ± 0.09) through sequence comparisons against a set of known type III effectors (Table 1). The Profile Kernel SVM is a *de novo* prediction approach that finds short motifs of consecutive residues in a database of proteins with known type III effector function annotation, *i.e.* it uses sequence similarity information that is not available directly from sequence comparisons. Applied to all protein sequences, the Profile Kernel SVM annotated 60 ± 11% type III effectors at 92 ± 8% accuracy (F1 = 0.73 ± 0.11). Our new method, pEffect, successfully combines the complementary homology-based and de novo predictions, reaching high levels of 87 ± 7% accuracy and 95 ± 5% coverage (F1 = 0.91 ± 0.08) and outperforming each of its individual components. In fact, pEffect is so good that about 80% of effector proteins in our Development set are predicted at 97% accuracy (RI > 80, Figure 1).

pEffect predicts from the entire sequence – a useful feature for metagenomic analyses

pEffect distinguishes type III effectors from other bacterial and eukaryotic proteins using the full length sequence of proteins. The detection of N-terminal signals, often used as the only source of evidence for predicting type III effectors computationally, presents a special problem for metagenomic data because of the erroneous gene predictions and potentially absent reads in contig assemblies [49]. To bypass the assembly errors in evaluating the presence of type III secretion activity in a particular metagenomic sample, it would be helpful to annotate as coming from effector sequences protein fragments translated directly from the DNA reads. pEffect's ability to distinguish effectors from these fragments can provide, for further experimental follow-up, a broad overview of interactions taking place in the sequenced microbiomes. Notably, for all fragment sets tested, pEffect performance was within the standard error of that achieved using full-length sequences (Figure 1). This result suggests that the features distinguishing type III effectors are present throughout the protein sequence and are not solely confined to the N-terminal region. Moreover, pEffect results can help establish the presence or absence of pathogenic organisms in a particular environment.

Gram-negative bacteria with full T3SS have the highest number of predicted effectors

The loss of type III secretion components in gram-negative bacteria is accompanied by the loss of effectors, indicating the lack of necessity to further diversify in the absence of the complete machinery (Figure 5C). This type of correlation between the completeness of T3SS and the number of effectors in gram-negative bacteria is not present for non-type III secreting gram-positive bacteria (Figure 5B) or archaea (Figure 5A).

Most of pEffect predictions are SVM-based

Type III effectors were predicted in all types of prokaryotes that we tested. As expected, the number of effectors in gram-positive bacteria and archaea that are not known to utilize T3SS was lower than in gram-negative bacteria that do use the system (Figures 4-5). Interestingly, homology searches, *i.e.* PSI-BLAST results, have identified roughly equal numbers of effectors (1%; Figure 6) in both types of bacterial genomes. As some effectors often co-localize with the T3SS machinery in “pathogenicity islands” [50-52], these findings are in line with the inheritance of the early complete secretory system, including the machinery and the secreted proteins.

Overall, the percentage of by similarity predicted effectors ranged for bacteria between 3% and 71% (maximum in *Onion yellows phytoplasma* OY-M, an intracellular gram-negative plant pathogen [8]), and averaged at 1%. Conversely, a significantly larger fraction, on average ~76% of all effector predictions in whole proteomes, was made *de novo*. The percentage of *de novo* predictions in gram-negative bacteria was significantly larger than in gram-positive ones (79 ± 0.4% vs. 70 ± 0.5%, respectively; Figure 4). Note, however, that 70% is still a drastically large fraction to appear in bacteria that seemingly have no use for them. Furthermore, the number of “new” (*i.e. de novo*) effectors has grown over evolutionary time (Figure 7), suggesting functional innovation due to environmental pressures. The set of *de novo* identified effectors found across bacteria is thus a good starting point for further investigation into effector origins.

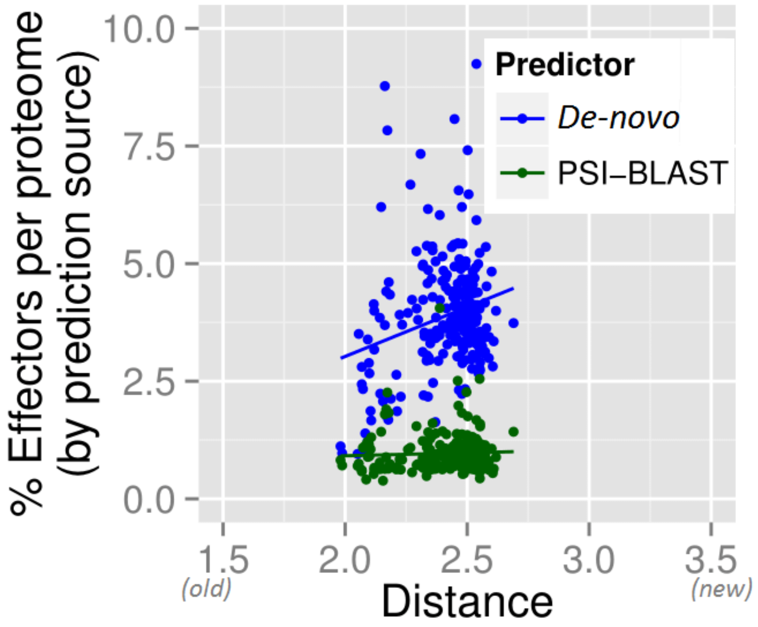


Figure 6: pEffect's whole proteome predictions identified by source. pEffect predicted type III effector proteins in the proteomes of 294 gram-negative and 29 gram-positive bacteria having full T3SS. The proteomes are shown as dots. Green dots indicate the percentage of proteins predicted as effectors (Y-axis) by homology searches and blue dots are *de novo* predictions. For each proteome, the evolutionary distance from the last common ancestor (X-axis) was extracted from [36]. While PSI-BLAST appears to consistently pick up ~1% of each proteome of all organisms (green horizontal trend-line), the *de novo* predicted effectors diversify further over evolutionary distance, as indicated by the increase in the number of *de novo* predictions.

Further insight into evolution of bacterial T3SS

pEffect's high prediction accuracy raises an interesting question about its predictions of effectors in gram-positive bacteria, which are not known to utilize T3SS. Roughly one fourth of their predicted effectors are of yet unknown function (Table 2). Bacterial proteins of annotated function are mostly transferases, hydrolases, ATP-binding proteins or kinases, all of which are necessary for flagellar motility. This finding is in line with evidence of shared ancestry between bacterial flagellar and type III secretion systems [35]. It is not known whether T3SS evolved from the flagellar apparatus or if the two systems evolved in parallel. However, gene genealogies [53] and protein network analysis approaches [54] both suggest independent evolution from a common ancestor, which comprised a subset of proteins forming a membrane-bound complex. The fact that the flagellar system can also secrete proteins [55] suggests that this ancestor may have played a secretory role [35]. The pervasiveness of the flagellar apparatus across the bacterial space suggests that the ancestral complex existed prior to the split of the cell-walled and double-membrane organisms, indicated by the differences in gram staining. The common ancestor protein complex of T3SS and flagellar system would have then been encoded in an even earlier ancestral genome. Thus, it is not surprising that we find T3SS component homology in gram-positive bacteria even in the absence of type III secretion functionality. Interestingly, our results show that the loss of the complete T3SS and, inherently, the associated loss in type III functionality has proceeded at a roughly similar rate in gram-positive and gram-negative bacteria (Figure 7A); *i.e.* once the T3SS is incomplete (4 components), and arguably non-functional, further loss of components consistently follows. A complete T3SS, however, is only visible in early gram-positive bacteria, but preserved across time in gram-negative bacteria (Figure 7B), further confirming the presence of the ancestral secretory complex in the last common bacterial ancestor.

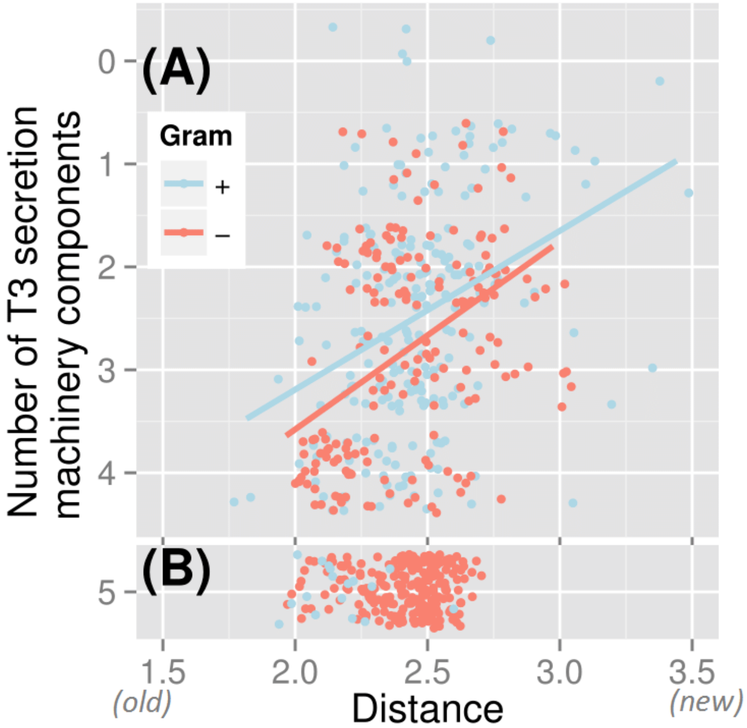


Figure 7: Loss of T3SS functionality differentiates gram-positive and gram-negative bacteria. 274 gram-positive bacteria (blue dots) and 588 gram-negative bacteria (red dots) were screened for the number of conserved components of T3SS (max. 5 T3DB Ortholog clusters) in their genomes (Y-axis) and plotted against the evolutionary distance from the most recent common ancestor (X-axis). **(A)** Once the T3SS is lost, *i.e.* less than five components are present, further rate of loss of components is the same for all bacteria. **(B)** The number of gram-negative bacteria with the complete system, *i.e.* all five components are present, however, remains constant across evolutionary time, while the number of gram-positive bacteria declines.

Did T3SS exist before the split of archaea and bacteria?

pEffect predicts a significant number of effectors in archaea. However, the presence of the beginnings of T3SS in the common ancestor of bacteria and archaea is neither directly supported nor negated by our results. Archaeal flagella have little or no structural similarities to bacterial flagella, but share homology with the bacterial type IV secretion system [56]. Some of the type IV secretion system and T3SS components are homologous, *e.g.* VirB11-like ATPases [57]. However, despite this observed homology none of the archaea that we tested had the complete set of T3SS components (Figure 5). If the common ancestor of archaea and bacteria did encode the core ancestral complex, these observations would indicate a loss of functionality in archaea. Another possibility is that the T3SS in bacteria, like the flagellar apparatus [58], may have been built over time from duplicated and diversified paralogous genes of the core complex after the archaea/bacteria split. In both of these scenarios, the prediction of type III effectors in archaea would then indicate re-purposing of the proteins secreted by the ancestral complex. In fact, 0.5% of an average archaeal genome is identified by homology (PSI-BLAST) to known effectors and another 0.9% *de novo* identified proteins are homologous (PSI-BLAST E-value ≤ 10^{-3}) to predicted effectors of gram-negative bacteria. These proteins must have been re-purposed in modern archaea; they are usually annotated as hydrolases, transferases, and metal-binding proteins (Table 2). The use of an additional 0.5% of the archaeal proteome that is picked up by pEffect's *de novo* and has no homologs in bacteria remains an enigma. While a certain level of similarity exists between archaeal proteins and bacterial type III effectors machinery, the observed signal is insufficient to draw definitive conclusions regarding common ancestry. It is, however, significant for further exploration – if roughly one tenth of the identified effectors of gram-negative bacteria and half of the machinery have homologs in archaea, could there have been a common ancestral secretion complex that has developed early on in evolutionary time and has given root to many systems observed today?

6.5 References

1. Dean P: **Functional domains and motifs of bacterial type III effector proteins and their roles in infection.** *FEMS microbiology reviews* 2011, **35:**1100-1125.
2. Cornelis GR: **The type III secretion injectisome.** *Nature reviews Microbiology* 2006, **4:**811-825.
3. Macho AP: **Subversion of plant cellular functions by bacterial type-III effectors: beyond suppression of immunity.** *The New phytologist* 2015.
4. Ferrell JC, Fields KA: **A working model for the type III secretion mechanism in Chlamydia.** *Microbes and infection / Institut Pasteur* 2015.
5. Dale C, Plague GR, Wang B, Ochman H, Moran NA: **Type III secretion systems and the evolution of mutualistic endosymbiosis.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99:**12397-12402.
6. Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T: **Sequence-based prediction of type III secreted proteins.** *PLoS pathogens* 2009, **5:**e1000376.
7. Wang Y, Zhang Q, Sun MA, Guo D: **High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles.** *Bioinformatics* 2011, **27:**777-784.
8. Lower M, Schneider G: **Prediction of type III secretion signals in genomes of gram-negative bacteria.** *PloS one* 2009, **4:**e5917.
9. Wang Y, Sun M, Bao H, White AP: **T3_MM: A Markov Model Effectively Classifies Bacterial Type III Secretion Signals.** *PloS one* 2013, **8:**e58173.
10. McDermott JE, Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne ED, Sharp D, Adkins JN, Samudrala R, Heffron F: **Computational prediction of type III and IV secreted effectors in gram-negative bacteria.** *Infection and immunity* 2011, **79:**23-32.
11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic acids research* 1997, **25:**3389-3402.
12. Cortes C, Vapnik V: **Support-Vector Networks.** *Machine Learning* 1995, **20:**273-297.
13. Kuang R, Ie E, Wang K, Siddiqi M, Freund Y, Leslie C: **Profile-based string kernels for remote homology detection and motif extraction.** *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB IEEE Computational Systems Bioinformatics Conference* 2004**:**152-160.
14. Hamp T, Goldberg T, Rost B: **Accelerating the Original Profile Kernel.** *PloS one* 2013, **8:**e68459.
15. UniProt Consortum: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** *Nucleic acids research* 2012, **40:**D71-75.
16. Angot A, Vergunst A, Genin S, Peeters N: **Exploitation of eukaryotic ubiquitin signaling pathways by effectors translocated by bacterial type III and type IV secretion systems.** *PLoS pathogens* 2007, **3:**e3.
17. Chang JH, Urbach JM, Law TF, Arnold LW, Hu A, Gombar S, Grant SR, Ausubel FM, Dangl JL: **A high-throughput, near-saturating screen for type III effector genes from Pseudomonas syringae.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:**2549-2554.
18. Greenberg JT, Vinatzer BA: **Identifying type III effectors of plant pathogens and analyzing their interaction with plant cells.** *Current opinion in microbiology* 2003, **6:**20-28.

19. Gurlebeck D, Thieme F, Bonas U: **Type III effector proteins from the plant pathogen Xanthomonas and their role in the interaction with the host plant.** *Journal of plant physiology* 2006, **163:**233-255.

20. Guttman DS, Vinatzer BA, Sarkar SF, Ranall MV, Kettler G, Greenberg JT: **A functional screen for the type III (Hrp) secretome of the plant pathogen Pseudomonas syringae.** *Science* 2002, **295:**1722-1726.

21. Miao EA, Miller SI: **A conserved amino acid sequence directing intracellular type III secretion by Salmonella typhimurium.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97:**7539-7544.

22. Sato H, Frank DW: **ExoU is a potent intracellular phospholipase.** *Molecular microbiology* 2004, **53:**1279-1290.

23. Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, Fivian A, Younis R, Matthews S, Marches O, Frankel G, et al: **An extensive repertoire of type III secretion effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103:**14941-14946.

24. Lindeberg M: **http://www.pseudomonas-syringae.org/**.

25. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic acids research* 2000, **28:**45-48.

26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215:**403-410.

27. Mika S, Rost B: **UniqueProt: Creating representative protein sequence sets.** *Nucleic acids research* 2003, **31:**3789-3791.

28. Rost B: **Twilight zone of protein sequence alignments.** *Protein engineering* 1999, **12:**85-94.

29. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9:**56-68.

30. Wang Y, Huang H, Sun M, Zhang Q, Guo D: **T3DB: an integrated database for bacterial type III secretion system.** *BMC bioinformatics* 2012, **13:**66.

31. Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, Altermann U, Angerer P, Ansorge S, Balasz K, et al: **LocTree3 prediction of localization.** *Nucleic acids research* 2014, **42:**W350-355.

32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic acids research* 2000, **28:**235-242.

33. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka.** *Bioinformatics* 2004, **20:**2479-2481.

34. http://biocomputer.bio.cuhk.edu.hk/T3DB/T3-ortholog-clusters.php.

35. McCann HC, Guttman DS: **Evolution of the type III secretion system and its effectors in plant-microbe interactions.** *The New phytologist* 2008, **177:**33-47.

36. Lang JM, Darling AE, Eisen JA: **Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices.** *PloS one* 2013, **8:**e62510.

37. Zhou Q, Su X, Ning K: **Assessment of quality control approaches for metagenomic data analysis.** *Scientific reports* 2014, **4:**6957.

38. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends in genetics : TIG* 2000, **16:**276-277.

39. http://www.ebi.ac.uk/genomes.

40. Konovalova A, Petters T, Sogaard-Andersen L: **Extracellular biology of Myxococcus xanthus.** *FEMS microbiology reviews* 2010, **34:**89-106.

41. Attree O, Attree I: **A second type III secretion system in Burkholderia pseudomallei: who is the real culprit?** *Microbiology* 2001, **147:**3197-3199.

42. Bertelli C, Collyn F, Croxatto A, Ruckert C, Polkinghorne A, Kebbi-Beghdadi C, Goesmann A, Vaughan L, Greub G: **The Waddlia genome: a window into chlamydial biology.** *PloS one* 2010, **5:**e10890.
43. Block A, Guo M, Li G, Elowsky C, Clemente TE, Alfano JR: **The Pseudomonas syringae type III effector HopG1 targets mitochondria, alters plant development and suppresses plant innate immunity.** *Cellular microbiology* 2010, **12:**318-330.
44. Brugirard-Ricaud K, Givaudan A, Parkhill J, Boemare N, Kunst F, Zumbihl R, Duchaud E: **Variation in the effectors of the type III secretion system among Photorhabdus species as revealed by genomic analysis.** *Journal of bacteriology* 2004, **186:**4376-4381.
45. Dai W, Li Z: **Conserved type III secretion system exerts important roles in Chlamydia trachomatis.** *International journal of clinical and experimental pathology* 2014, **7:**5404-5414.
46. Mavrodi DV, Joe A, Mavrodi OV, Hassan KA, Weller DM, Paulsen IT, Loper JE, Alfano JR, Thomashow LS: **Structural and functional analysis of the type III secretion system from Pseudomonas fluorescens Q8r1-96.** *Journal of bacteriology* 2011, **193:**177-189.
47. Salinero KK, Keller K, Feil WS, Feil H, Trong S, Di Bartolo G, Lapidus A: **Metabolic analysis of the soil microbe Dechloromonas aromatica str. RCB: indications of a surprisingly complex life-style and cryptic anaerobic pathways for aromatic degradation.** *BMC genomics* 2009, **10:**351.
48. Przybylski D, Rost B: **Alignments grow, secondary structure prediction improves.** *Proteins* 2002, **46:**197-205.
49. Nielsen P, Krogh A: **Large-scale prokaryotic gene prediction and comparison to genome annotation.** *Bioinformatics* 2005, **21:**4322-4329.
50. Figueira R, Holden DW: **Functions of the Salmonella pathogenicity island 2 (SPI-2) type III secretion system effectors.** *Microbiology* 2012, **158:**1147-1161.
51. Okada N, Iida T, Park KS, Goto N, Yasunaga T, Hiyoshi H, Matsuda S, Kodama T, Honda T: **Identification and characterization of a novel type III secretion system in trh-positive Vibrio parahaemolyticus strain TH3996 reveal genetic lineage and diversity of pathogenic machinery beyond the species level.** *Infection and immunity* 2009, **77:**904-913.
52. Reis RS, Horn F: **Enteropathogenic Escherichia coli, Samonella, Shigella and Yersinia: cellular aspects of host-bacteria interactions in enteric diseases.** *Gut pathogens* 2010, **2:**8.
53. Gophna U, Ron EZ, Graur D: **Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events.** *Gene* 2003, **312:**151-163.
54. Medini D, Covacci A, Donati C: **Protein homology network families reveal step-wise diversification of Type III and Type IV secretion systems.** *PLoS computational biology* 2006, **2:**e173.
55. Macnab RM: **Type III flagellar protein export and flagellar assembly.** *Biochimica et biophysica acta* 2004, **1694:**207-217.
56. Ng SY, Chaban B, Jarrell KF: **Archaeal flagella, bacterial flagella and type IV pili: a comparison of genes and posttranslational modifications.** *Journal of molecular microbiology and biotechnology* 2006, **11:**167-191.
57. Wallden K, Rivera-Calzada A, Waksman G: **Type IV secretion systems: versatility and diversity in function.** *Cellular microbiology* 2010, **12:**1203-1212.
58. Liu R, Ochman H: **Stepwise formation of the bacterial flagellar system.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104:**7116-7121.

7 LocText: a manually annotated text corpus for protein localization data

7.1 Preface

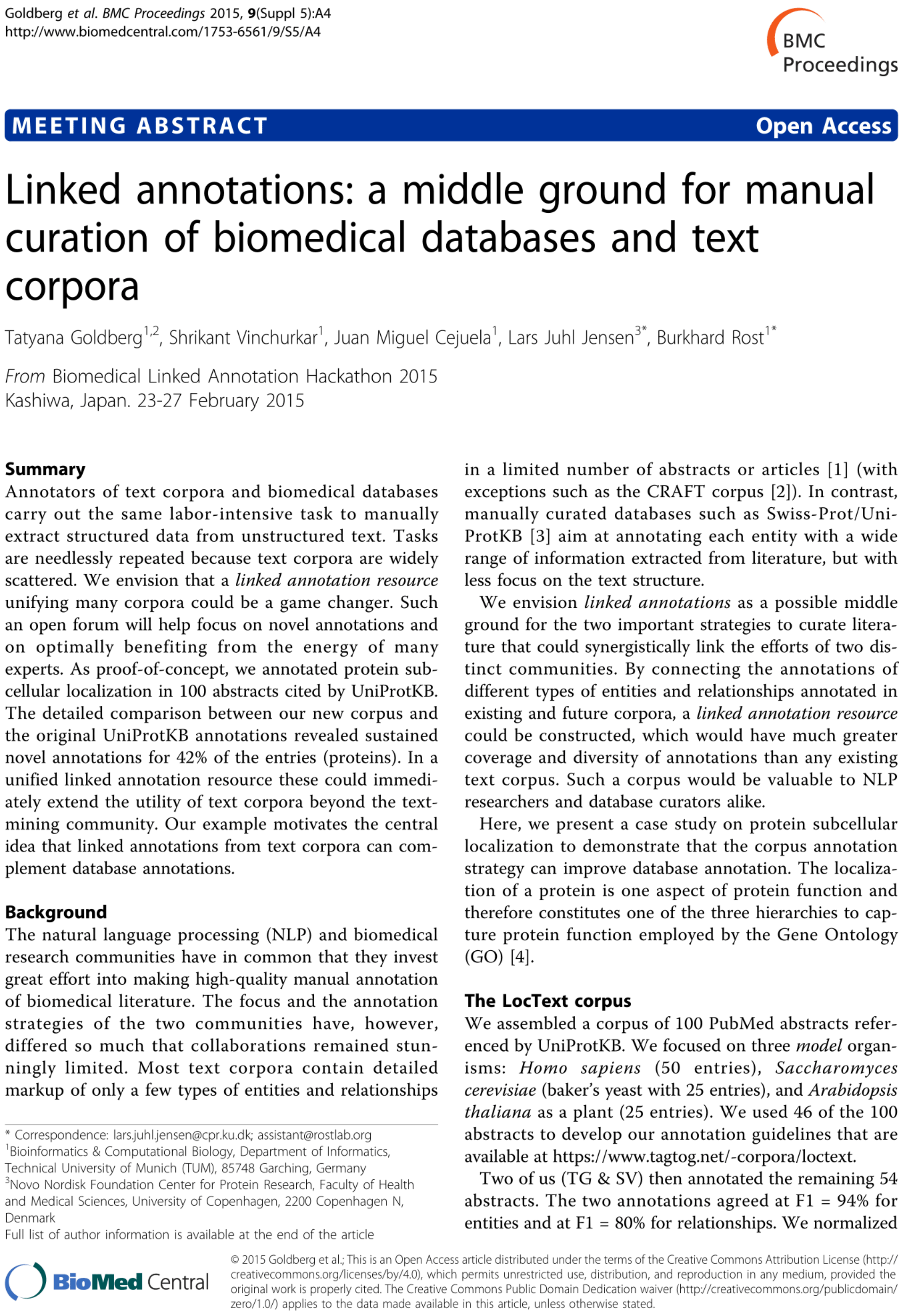
Scientific literature is the central repository for scientific knowledge. Having access to this accumulated knowledge enables researched to efficiently generate novel knowledge. For example, PubMed [1] is a widely used database [2] that stores over 25 million records for biomedical literature; 500,000 new records are added to the database each year [3]. At this high rate of knowledge extension, it is impossible to manually extract structured data (*e.g.* aspects of protein structure and function) from unstructured texts (*i.e.* literature records).

Many databases have been developed to store structured data from scientific publications and to make it instantly available for researchers online. In the area of life sciences, the most prominent examples are UniProt [4], GenBank [5], Ensembl [6] and others. Another resource that stores structured data from scientific publications is presented by text corpora that are developed to train machine learning methods for automated text recognition [7-9]. While database curators aim to annotate a single entity (usually a gene or protein) with a wide range of information extracted from literature, the curators of text corpora focus on a detailed markup of only a few entities and relationships in a limited number of literature records. Because of the different focus and the annotation strategies of the two communities, collaborations between them remained stunningly limited.

In this publication, we envision a *linked annotation resource* unifying many corpora and database entries to be a game changer. By connecting the annotations of different types of entities, a linked resource could have a much greater coverage and diversity than any single resource. As proof-of-concept, we annotated protein sub-cellular localization in 100 abstracts cited by UniProt. By comparing our new corpus with the original UniProt annotations, we found novel annotations for 42% of the protein entries. Thus, we showed that a linked resource could complement database annotations with those from text corpora.

The study design was conceived by me, Juan Miguel Cejuela and Lars Juhl Jensen. Abstract annotations were done by me, Juan Miguel Cejuela and Shrikant Vinchurkar. All calculations were done by me with the help of Lars Juhl Jensen. The manuscript was drafted by me, Lars Juhl Jensen and Burkhard Rost.

7.2 Journal article. Goldberg T., Vinchurkar S., Cejuela J.M., Jensen L.J. , Rost B. *BMC Proceedings* 2015, 9(Suppl 5):A4



Goldberg *et al. BMC Proceedings* 2015, **9**(Suppl 5):A4
http://www.biomedcentral.com/1753-6561/9/S5/A4

BMC Proceedings

MEETING ABSTRACT Open Access

Linked annotations: a middle ground for manual curation of biomedical databases and text corpora

Tatyana Goldberg[1,2], Shrikant Vinchurkar[1], Juan Miguel Cejuela[1], Lars Juhl Jensen[3*], Burkhard Rost[1*]

From Biomedical Linked Annotation Hackathon 2015
Kashiwa, Japan. 23-27 February 2015

Summary

Annotators of text corpora and biomedical databases carry out the same labor-intensive task to manually extract structured data from unstructured text. Tasks are needlessly repeated because text corpora are widely scattered. We envision that a *linked annotation resource* unifying many corpora could be a game changer. Such an open forum will help focus on novel annotations and on optimally benefiting from the energy of many experts. As proof-of-concept, we annotated protein subcellular localization in 100 abstracts cited by UniProtKB. The detailed comparison between our new corpus and the original UniProtKB annotations revealed sustained novel annotations for 42% of the entries (proteins). In a unified linked annotation resource these could immediately extend the utility of text corpora beyond the text-mining community. Our example motivates the central idea that linked annotations from text corpora can complement database annotations.

Background

The natural language processing (NLP) and biomedical research communities have in common that they invest great effort into making high-quality manual annotation of biomedical literature. The focus and the annotation strategies of the two communities have, however, differed so much that collaborations remained stunningly limited. Most text corpora contain detailed markup of only a few types of entities and relationships in a limited number of abstracts or articles [1] (with exceptions such as the CRAFT corpus [2]). In contrast, manually curated databases such as Swiss-Prot/UniProtKB [3] aim at annotating each entity with a wide range of information extracted from literature, but with less focus on the text structure.

We envision *linked annotations* as a possible middle ground for the two important strategies to curate literature that could synergistically link the efforts of two distinct communities. By connecting the annotations of different types of entities and relationships annotated in existing and future corpora, a *linked annotation resource* could be constructed, which would have much greater coverage and diversity of annotations than any existing text corpus. Such a corpus would be valuable to NLP researchers and database curators alike.

Here, we present a case study on protein subcellular localization to demonstrate that the corpus annotation strategy can improve database annotation. The localization of a protein is one aspect of protein function and therefore constitutes one of the three hierarchies to capture protein function employed by the Gene Ontology (GO) [4].

The LocText corpus

We assembled a corpus of 100 PubMed abstracts referenced by UniProtKB. We focused on three *model* organisms: *Homo sapiens* (50 entries), *Saccharomyces cerevisiae* (baker's yeast with 25 entries), and *Arabidopsis thaliana* as a plant (25 entries). We used 46 of the 100 abstracts to develop our annotation guidelines that are available at https://www.tagtog.net/-corpora/loctext.

Two of us (TG & SV) then annotated the remaining 54 abstracts. The two annotations agreed at F1 = 94% for entities and at F1 = 80% for relationships. We normalized

* Correspondence: lars.juhl.jensen@cpr.ku.dk; assistant@rostlab.org
[1] Bioinformatics & Computational Biology, Department of Informatics, Technical University of Munich (TUM), 85748 Garching, Germany
[3] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark
Full list of author information is available at the end of the article

protein names to UniProtKB and localizations to GO identifiers. The resulting corpus contains 306 annotated relationships in 201 different UniProtKB proteins with 48 GO distinct localization terms. All annotations were made within the framework of the *tagtog* system (Figure 1; http://tagtog.net) [5] and Reflect was used to aid protein name normalization (http://reflect.ws) [6]. The corpus is available for download at https://www.tagtog.net/-corpora/loctext under the Creative Commons Attribution 4.0 (CC-BY 4.0) license.

Corpus provides novel annotations

Linked annotations from text corpora can complement database annotations only if manual corpus annotations identify relationships not captured by existing databases. Therefore, all our annotations were done from scratch without using database annotations. Comparing our "from scratch" annotations with those from UniProtKB revealed important novelty added by our text corpus.

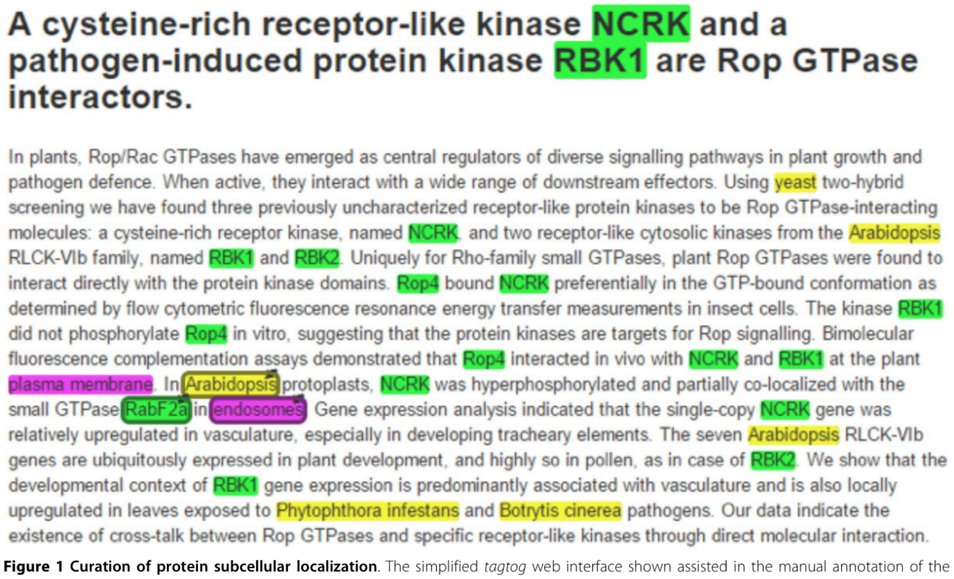
We found novel or more detailed localization annotations with respect to UniProtKB for 84 of 201 (42%) proteins in 34 abstracts (Table 1); for example, *Arabidopsis* RabF2a (UniProtKB entry RAF2A_ARATH) is localized to endosomes (Figure 1). We found that for over half of these proteins with additional annotations (47/84 = 56%) UniProtKB did not cite the abstracts. This is likely explained by the way proteins are annotated, one protein at a time: if a curator works on one protein and an abstract mentions also the localization of another, which is not the focus of curator, the localization of the latter might not be annotated.

Table 1. Localization annotations in our corpus and in UniProtKB. The table categorizes the corpus relationships by organism relative to whether they represent existing annotations in UniProtKB, more detailed annotations, or truly novel annotations. It further subdivides the counts based on whether or not the relationships involve UniProtKB proteins that cite the abstract

Category	Existing		More detailed		Novel	
Citing protein	Yes	No	Yes	No	Yes	No
Human	29	15	1	1	14	13
Budding yeast	22	14	5	3	6	15
Arabidopsis	19	7	5	2	6	7
Other	2	9	0	0	0	6
Subtotal	72	45	11	6	26	41
Total	117		17		67	

Perspectives

Our case study clearly showed that corpora containing manual annotations of the sub-cellular localization of



A cysteine-rich receptor-like kinase NCRK and a pathogen-induced protein kinase RBK1 are Rop GTPase interactors.

In plants, Rop/Rac GTPases have emerged as central regulators of diverse signalling pathways in plant growth and pathogen defence. When active, they interact with a wide range of downstream effectors. Using yeast two-hybrid screening we have found three previously uncharacterized receptor-like protein kinases to be Rop GTPase-interacting molecules: a cysteine-rich receptor kinase, named NCRK, and two receptor-like cytosolic kinases from the Arabidopsis RLCK-VIb family, named RBK1 and RBK2. Uniquely for Rho-family small GTPases, plant Rop GTPases were found to interact directly with the protein kinase domains. Rop4 bound NCRK preferentially in the GTP-bound conformation as determined by flow cytometric fluorescence resonance energy transfer measurements in insect cells. The kinase RBK1 did not phosphorylate Rop4 in vitro, suggesting that the protein kinases are targets for Rop signalling. Bimolecular fluorescence complementation assays demonstrated that Rop4 interacted in vivo with NCRK and RBK1 at the plant plasma membrane. In Arabidopsis protoplasts, NCRK was hyperphosphorylated and partially co-localized with the small GTPase RabF2a in endosomes. Gene expression analysis indicated that the single-copy NCRK gene was relatively upregulated in vasculature, especially in developing tracheary elements. The seven Arabidopsis RLCK-VIb genes are ubiquitously expressed in plant development, and highly so in pollen, as in case of RBK2. We show that the developmental context of RBK1 gene expression is predominantly associated with vasculature and is also locally upregulated in leaves exposed to Phytophthora infestans and Botrytis cinerea pathogens. Our data indicate the existence of cross-talk between Rop GTPases and specific receptor-like kinases through direct molecular interaction.

Figure 1 Curation of protein subcellular localization. The simplified *tagtog* web interface shown assisted in the manual annotation of the corpus (abstract of [7]). Colours highlight names of organisms (yellow), genes/proteins (green), and localization terms (magenta). Linking the *Arabidopsis* protein RabF2a (UniProtKB ID: RAF2A_ARATH) to endosomes adds a novel annotation to UniProtKB.

proteins are able to contribute novel information to curated databases such as UniProtKB. Notably, this is even true in the worst-case example when limiting annotations only to abstracts of articles that have already been utilized by the database curators. We expect our findings to generalize to most types of protein annotation, including disease associations and tissue expression.

Today databases avoid the trouble of integrating these annotations, because most text corpora are too limited in size and scope. Having the corpus developers combine their annotations into a single, unified linked annotation resource could thus be an important step towards integration of corpus annotations into databases, thus making them to richer data collection systems. Even before integration with databases happens, it will be possible for researchers to use semantic web technologies to combine the information in the linked annotation resource with that in existing databases, since UniProtKB and many other databases are already Resource Description Framework (RDF) compliant.

We envision a linked annotation resource to continuously grow, supported by annotation tools making it easy for corpus developers to link future annotations; for example, through a standard JSON format. Not all linked annotations need to be made manually, though. Including also results from automatic text mining pipelines would help address the challenge of the prohibitively high costs of large-scale manual annotation [2]. Associations extracted from both open and non-open access journals can be linked, as redistribution of extracted facts is not prohibited by most publishers' licenses.

Acknowledgements
Funding: Alexander von Humboldt Foundation through German Federal Ministry for Education and Research, Ernst Ludwig Ehrlich Studienwerk, and the Novo Nordisk Foundation Center for Protein Research (NNF14CC0001).

Authors' details
[1]Bioinformatics & Computational Biology, Department of Informatics, Technical University of Munich (TUM), 85748 Garching, Germany. [2]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), 85748 Garching, Germany. [3]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark.

Published: 6 August 2015

References

1. Neves M: **An analysis on the entity annotations in biological corpora.** *F1000Res* 2014, **3**:96.
2. Verspoor K, Cohen KB, Lanfranchi A, Warner C, Johnson HL, Roeder C, Choi JD, Funk C, Malenkiy Y, Eckert M, Xue N, Baumgartner WA Jr, Bada M, Palmer M, Hunter LE: **A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools.** *BMC Bioinformatics* 2012, **13**:207.
3. UniProt Consortium: **Activities at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2014, **42**:D191-D198.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.
5. Cejuela JM, McQuilton P, Ponting L, Marygold SJ, Stefancsik R, Millburn GH, Rost B, FlyBase Consortium: **tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles.** *Database* 2014, **2014**:bau033.
6. Pafilis E, O'Donoghue SI, Jensen LJ, Horn H, Kuhn M, Brown NP, Schneider R: **Reflect: augmented browsing for the life scientist.** *Nat Biotechnol* 2009, **27**:508-510.
7. Molendijk AJ, Ruperti B, Singh MK, Dovzhenko A, Ditengou FA, Milia M, Westphal L, Rosahl S, Soellick TR, Uhrig J, Weingarten L, Huber M, Palme K: **A cysteine-rich receptor-like kinase NCRK and a pathogen-induced protein kinase RBK1 are Rop GTPase interactors.** *Plant J* 2008, **53**:909-923.
8. Baumgartner WA, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L: **Manual curation is not sufficient for annotation of genomic databases.** *Bioinformatics* 2007, **23(13)**:i41-i48.

doi:10.1186/1753-6561-9-S5-A4
Cite this article as: Goldberg *et al.*: Linked annotations: a middle ground for manual curation of biomedical databases and text corpora. *BMC Proceedings* 2015 **9**(Suppl 5):A4.

7.3 References

1. http://www.ncbi.nlm.nih.gov/pubmed.
2. Yoo I, Mosa AS: **Analysis of PubMed User Sessions Using a Full-Day PubMed Query Log: A Comparison of Experienced and Nonexperienced PubMed Users.** *JMIR medical informatics* 2015, **3:**e25.
3. Lu Z: **PubMed and beyond: a survey of web tools for searching biomedical literature.** *Database : the journal of biological databases and curation* 2011, **2011:**baq036.
4. **UniProt: a hub for protein information.** *Nucleic acids research* 2015, **43:**D204-212.
5. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic acids research* 2009, **37:**D26-31.
6. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al: **Ensembl 2015.** *Nucleic acids research* 2015, **43:**D662-669.
7. Neves M: **An analysis on the entity annotations in biological corpora.** *F1000Research* 2014, **3:**96.
8. Van Auken K, Schaeffer ML, McQuilton P, Laulederkind SJ, Li D, Wang SJ, Hayman GT, Tweedie S, Arighi CN, Done J, et al: **BC4GO: a full-text corpus for the BioCreative IV GO task.** *Database : the journal of biological databases and curation* 2014, **2014**.
9. Pyysalo S, Ohta T, Miwa M, Cho HC, Tsujii J, Ananiadou S: **Event extraction across multiple levels of biological organization.** *Bioinformatics* 2012, **28:**i575-i581.

8 Conclusions

8.1 Our work in the context of developments in the field

In silico prediction of protein cellular sorting is one of the main testing grounds for the development of prediction methods for protein function. Over the last two decades, more experimental data for protein localization became available, and many methods have been developed to predict protein localization. The methods apply various algorithms for their predictions. The widely used methods are those that apply machine learning (ML) techniques to extract information encoded in the amino acid sequences of proteins.

In our work, we built upon the experience of previously published methods and developed a novel method for sub-cellular localization prediction [1]. We employed Support Vector Machines (SVM) [1], an ML technique that was previously shown to perform best for localization predictions [2]. We implemented SVMs in a hierarchical tree to mimic the protein sorting mechanism, an idea originally introduced by Nair and Rost [3]. However, we ignored many of the relevant features used for the success of other methods (*e.g.* we ignored aspects of protein structure and function [3, 4], signal peptides [3] and other functional motifs [5-7], and physicochemical properties of amino acids [7-9]). Instead, we used advanced SVM Profile Kernels [10, 11] that at all levels of the tree search through proteins of annotated localization with short stretches of *k*-consecutive residues (k=6 for eukaryota, 5 for bacteria and 3 for archaea) and match those in a query protein. The most informative *k*-mer hit then decides on the “left or right” at each decision point in the tree until reaching a leaf, *i.e.* the predicted localization class. Thus, SVMs reach their predictions through levels of sequence similarity that are not available directly through sequence comparisons.

The novel method LocTree2 predicted protein localization in all domains of life in the so far largest number of protein localization compartments (18 classes for eukaryota, 6 for bacteria and 3 for archaea). It outperformed other methods, including experts specialized in distinguishing between proteins of two classes [12, 13], implicating an improved ability of our method to capture localization signals in the protein sequence. Another important improvement was the robustness of the method against sequencing errors and its success when applied to protein fragments. This is particularly important in light of high-throughput sequencing, of analyzing ancient DNA with short reads and of the fact that almost 80% of all proteins have multiple domains [14].

We could further improve LocTree2 by remarkable 25% by including information about homologs if available. These were obtained through PSI-BLAST [15] searches. PSI-BLAST has certainly changed the way we do sequence analysis more than any tool and it has been continuously improving since its publication in 1997 adding important value beyond that from growing databases [16]. We found that in development set of LocTree2, about half of all proteins have experimentally annotated homologs. For these proteins, a simple PSI-BLAST protocol significantly outperformed LocTree2, which is in line with the findings of Imai and Nakai [17]. For the other half of proteins, the homology-based inference became random, dropping the performance significantly below that of LocTree2. Our new method, LocTree3, successfully combined homology-based and *de novo* predictions of localization, reaching an 18-state accuracy Q18 = 80 ± 3% for eukaryotes and a six-state accuracy Q6 = 89 ± 4% for bacteria. We made the method publicly available as a web server, allowing submissions to range from single protein sequences to entire proteomes. Due to its high prediction performance, short prediction time and cached results, LocTree3 optimized well for the handling of large-scale data and aiding the prediction of protein function through localization predictions.

The prediction results of LocTree2 and LocTree3 have already been found useful for complementing experimental annotations and for improving protein function prediction methods. For example, both methods were cited for identifying proteins of the human multicellular signaling network [18], improving predictions of protein-protein interactions [19, 20], determining cell surface proteins of the human immune system [21], localizing proteins of human cancer cells [22], identifying plant pathogens [23] and characterizing proteins that improve resistance in plants [24-26]. We believe that the framework for our methods will prove extendable and that future methods will become better simply by using more experimental data and more sequences.

The success of LocTree3's approach – use homology information if available and a *de novo* prediction otherwise – has proven to hold true also for other classification problems, such as the prediction of sub-nuclear localization compartments and of bacterial pathogens. In the following, I will summarize some of the main findings of our research.

8.2 Summary of our main findings

LocTree2: Highest performance due to improved underlying method

We rigorously benchmarked the prediction performance of LocTree2 to a number of state-of-the-art methods using several independent data sets. LocTree was one of the benchmarked methods; it originally introduced the hierarchical system of SVMs that resembles cellular sorting. LocTree2 outperformed LocTree on all data sets tested. Even when trained and tested on LocTree's development data (3 localization classes for bacteria and 6 classes for eukaryota), we observed LocTree2's overall prediction accuracy to (i) stay within the standard error of what was achieved on LocTree2's development set (6 years older set) and (ii) increase by 18% for bacteria and by 7% for eukaryota. Thus, the improvement of LocTree2 originated mainly from the underlying method advancement and not the increased training data set.

In silico predictions reveal problems of high-throughput experiments

LocDB is a database collecting localization annotations mostly from high-throughput experiments [27]. We compared the prediction performance of LocTree2 and of other methods using sequence-unique sets (sequence-unique with respect to all proteins in the set and to the training set of all methods tested) on LocDB proteins. We found all methods to perform substantially worse on LocDB data than on sequence-unique proteins from Swiss-Prot [28], whose localization annotations are mostly derived from low-throughput experiments. For example, on the *A. thaliana* set, LocTree2's performance decreased by 28% and WoLF PSORT's performance by 43%. How to interpret the data from LocDB?

As most annotations in LocDB originate from high-throughput experiments, it is very likely that LocDB contains proportionally more errors than Swiss-Prot, which might explain why all methods perform worse for the LocDB than for the Swiss-Prot data. On the other hand, we might also suspect that high-throughput experiments discover a reality invisible to traditional experimental methods and some of those invisible facts might reveal new sorting mechanisms. Such hidden mechanisms might or might not be 'discovered' by prediction methods. If not, those would explain many incorrect predictions.

Each prediction of LocTree2 is accompanied by a reliability index (RI) denoting the strength of a prediction (from unreliable RI=0 to highly trustable RI=100). Zooming into annotations of by LocTree2 misclassified proteins with a high reliability (RI>50), we found examples of proteins for which low-throughput annotations in literature contradicted high-throughput annotations in LocDB. Thus, the predictions judged as incorrect by LocDB but having very high LocTree2's RI scores indicated that the low performance inverts the real picture: rather LocDB annotations are wrong or ambiguous than the strong LocTree2 predictions. For a set of weakest LocTree2 predictions (RI<15), we observed the opposite.

Eukaryotic secreted and bacterial plasma membrane proteins predicted best

LocTree3 combined homology-based inferences of PSI-BLAST with *de novo* predictions of LocTree2. Assessed on a non-redundant data set, LocTree3 performed very well for archaeal proteins (three classes) with the overall level of accuracy suggested to reach 100%. This number is most likely an over-estimate due to the limited data. For bacteria (six classes), the overall accuracy was Q6 = 89 ± 4% and for eukaryota (18 classes) it was Q18 = 65 ± 3%. For bacteria, LocTree3 predicted best plasma membrane (accuracy: 96%, coverage: 95%) and cytoplasmic proteins (accuracy: 91%, coverage: 90%). For eukaryota, the best predicted class was secreted (accuracy: 88%, coverage: 96%), followed by nucleus (accuracy: 81%, coverage: 86%). While LocTree2 predicted classes with most experimental annotations best, we could not confirm the same trend for the PSI-BLAST protocol. Overall, our new method, LocTree3, still maintained a small correlation between performance and experimental annotations with respect to the compartments.

Multi-localized proteins difficult to assess

Studies have shown that up to one third of all proteins in a proteome are localized to more than one sub-cellular compartment [29-31]. Annotations of multi-localized proteins are also contained in Swiss-Prot. However, applying sequence redundancy reduction (through UniqueProt [32] at HSSP-value ≤ 0 [33, 34] to these proteins, their number dropped to 72 eukaryotic proteins. We applied LocTree3 to these proteins and considered the prediction correct if one of the experimentally observed classes had been predicted. Prediction result of Q18 = 65 ± 12% compared less favorably to Q18 = 80 ± 3% when assessed on single-localized proteins. This contradicted the intuition - picking one right from 18 is tougher than picking 2 and choosing the best-of-two. Why did performance drop on those proteins?

Our suspicion is that today's double annotations as a whole set are not good enough. We looked at LocTree3 predictions for five misclassified proteins with the highest RIs. One protein was uncharacterized, while for the remaining four we found experimental evidence for the predicted localization classes in other sources than Swiss-Prot. From these findings we concluded that the number of sequence-unique multi-localized proteins as we have them today in Swiss-Prot is rather small and the annotations of multi-localizations may be incomplete. Therefore, assessing prediction methods on these proteins may lead to underestimated results and incorrect implications.

Homology-based inferences not sufficient for whole proteome annotations

We annotated proteomes of more than 1,000 fully sequences organisms from all three domains of life with LocTree3. We observed that none of the proteomes could be fully annotated with homology searches (*i.e.* by PSI-BLAST). For example, for human, LocTree3 annotated remarkable 77% of the proteome through homology-based inference, of which 30% came from direct experimental annotations. For other organisms these numbers were lower. For yeast, LocTree3 annotated 68% of the proteome by PSI-BLAST, of which 51% were experimental annotations; for *A. thaliana* these numbers were 61% and 11%, respectively. For a prokaryote *A. pernix*, LocTree3 annotated only 8% of the proteome by PSI-BLAST; the remaining annotations came from its *de novo* component LocTreee2.

Q13 = 62% for predicting sub-nuclear compartments

Though sub-organellar compartments are difficult to predict due to sparse experimental data, LocNuclei adapted the prediction strategy of LocTree3 (combine homology information with *de novo* predictions) and classified sub-nuclear proteins in 13 classes at the high level of overall accuracy Q13 = 62 ± 3%. LocNuclei outperformed, the only during LocNuclei's development available other method for sub-nuclear localization prediction, NSort [35] (we re-trained LocNuclei on NSort's development data). We used LocNuclei to annotate the entire human nucleosome (6,230 proteins predicted as nuclear by LocTree3) and found 77% of all proteins to localize to the following four sub-nuclear compartments: nucleoplasm (30% of all annotations), chromatin (17%), nucleolus (17%) and PML bodies (13%). Adding in experimental protein-protein interaction data [36], we found most protein interaction pairs to occur within and between these four compartments. Interestingly, we found a high number of protein interactions also between perinucleolar proteins, composing <0.4% of all annotations in the human nucleosome, and proteins residing in the nucleoplasm, chromatin and nucleolus. Compared to proteins from other sub-cellular compartments, nuclear proteins tend to be most disordered. This feature allows nuclear proteins to diversify their functional roles, which is in line with experimental findings [37-39].

NLS can be mapped in 50% of nuclear proteins and NES in 29%

NLSdb [40] was the first database that attempted to collect known nuclear localization signals in a single resource. It also introduced the concept of "*in silico* mutagenesis" [41] that extended experimental signals by potential ones. Fifteen years later, we updated NLSdb with novel data, which now contains both nuclear localization signals (NLS) and nuclear export signals (NES). By doing so, we increased the number of by experts manually verified signals 28-fold and of potential signals 20-fold. Looking at the length distributions of verified NLS, we observed possible annotation mistakes for at least 20% of monopartite signals. While most of the signals in our verified set were of virus and human origin, we observed an enrichment of bipartite signals in plants and yeast. When clustered verified signals by sequence similarity, we identified consensus sequences for 40 clusters of monopartite NLS, 38 clusters of bipartite NLS, 5 clusters of PY-NLS and 27 clusters of NES. Currently, Swiss-Prot annotates 9% of nuclear proteins with NLS and 5% of nuclear proteins with NES. The original version of NLSdb increases the coverage for NLS to 19%, while the updated version increases the coverage to 50% for NLS and 29% for NES.

Bacterial type III secretion signal distributed over the entire protein sequence

The bacterial type III secretion system injects the so-called effector proteins directly into the cytoplasm of a host cell to promote infection. pEffect is a method that showed that a combination of homology searches and *de novo* predictions can successfully be applied to the prediction of effector proteins at 87% accuracy and 95% coverage. While other methods mainly employ information encoded in the N-terminal region of protein sequences for their predictions [42-44], pEffect uses information from the entire protein sequence. When compared to other methods on full length protein sequences, pEffect performed best on all data sets tested. Especially on data sets containing eukaryotic proteins, pEffect's exceeded by more than 0.58 in the F1 performance measure. When tested on sequence fragments similar in length to shotgun sequencing reads, pEffect's performance was not significantly different. These improvements are particularly important to *e.g.* annotate results from metagenomic studies. Moreover, they suggest that the features distinguishing type III effectors are spread over the entire protein sequence and are picked up by pEffect.

8.3 References

1. Goldberg T, Hamp T, Rost B: **LocTree2 predicts localization for all domains of life.** *Bioinformatics* 2012, **28:**i458-i465.
2. Hua S, Sun Z: **Support vector machine approach for protein subcellular localization prediction.** *Bioinformatics* 2001, **17:**721-728.
3. Nair R, Rost B: **Mimicking cellular sorting improves prediction of subcellular localization.** *Journal of molecular biology* 2005, **348:**85-100.
4. Blum T, Briesemeister S, Kohlbacher O: **MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction.** *BMC bioinformatics* 2009, **10:**274.
5. Nguyen Ba AN, Pogoutse A, Provart N, Moses AM: **NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction.** *BMC bioinformatics* 2009, **10:**202.
6. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26:**1608-1615.
7. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor.** *Nucleic acids research* 2007, **35:**W585-587.
8. Yu CS, Lin CJ, Hwang JK: **Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions.** *Protein science : a publication of the Protein Society* 2004, **13:**1402-1406.
9. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization.** *Proteins* 2006, **64:**643-651.
10. Kuang R, Ie E, Wang K, Siddiqi M, Freund Y, Leslie C: **Profile-based string kernels for remote homology detection and motif extraction.** *Journal of bioinformatics and computational biology* 2005, **3:**527-550.
11. Hamp T, Goldberg T, Rost B: **Accelerating the Original Profile Kernel.** *PloS one* 2013, **8:**e68459.
12. Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A: **Prediction of membrane-protein topology from first principles.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105:**7177-7181.
13. Kall L, Krogh A, Sonnhammer EL: **An HMM posterior decoder for sequence feature prediction that includes homology information.** *Bioinformatics* 2005, **21 Suppl 1:**i251-257.
14. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes.** *Journal of molecular biology* 2001, **310:**311-325.
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic acids research* 1997, **25:**3389-3402.
16. Przybylski D, Rost B: **Alignments grow, secondary structure prediction improves.** *Proteins* 2002, **46:**197-205.
17. Imai K, Nakai K: **Prediction of subcellular locations of proteins: where to proceed?** *Proteomics* 2010, **10:**3970-3983.
18. Ramilowski JA, Goldberg T, Harshbarger J, Kloppman E, Lizio M, Satagopam VP, Itoh M, Kawaji H, Carninci P, Rost B, Forrest AR: **A draft network of ligand-receptor-mediated multicellular signalling in human.** *Nature communications* 2015, **6:**7866.

19. Hamp T, Rost B: **Evolutionary profiles improve protein-protein interaction prediction from sequence.** *Bioinformatics* 2015, **31:**1945-1950.

20. Zahiri J, Mohammad-Noori M, Ebrahimpour R, Saadat S, Bozorgmehr JH, Goldberg T, Masoudi-Nejad A: **LocFuse: human protein-protein interaction prediction via classifier fusion using protein localization information.** *Genomics* 2014, **104:**496-503.

21. Graessel A, Hauck SM, von Toerne C, Kloppmann E, Goldberg T, Koppensteiner H, Schindler M, Knapp B, Krause L, Dietz K, et al: **A Combined Omics Approach to Generate the Surface Atlas of Human Naive CD4+ T Cells during Early T-Cell Receptor Activation.** *Molecular & cellular proteomics : MCP* 2015, **14:**2085-2102.

22. Wilz SW, Liu D, Liu C, Yang J: **Development of a test to identify bladder cancer in the urine of patients using mass spectroscopy and subcellular localization of the detected proteins.** *American journal of translational research* 2015, **7:**1458-1466.

23. Ranasinghe SL, Fischer K, Gobert GN, McManus DP: **A novel coagulation inhibitor from Schistosoma japonicum.** *Parasitology* 2015, **142:**1663-1672.

24. Guerra-Guimaraes L, Tenente R, Pinheiro C, Chaves I, Silva Mdo C, Cardoso FM, Planchon S, Barros DR, Renaut J, Ricardo CP: **Proteomic analysis of apoplastic fluid of Coffea arabica leaves highlights novel biomarkers for resistance against Hemileia vastatrix.** *Frontiers in plant science* 2015, **6:**478.

25. Prasanna VK, Venkatesh YP: **Characterization of onion lectin (Allium cepa agglutinin) as an immunomodulatory protein inducing Th1-type immune response in vitro.** *International immunopharmacology* 2015, **26:**304-313.

26. Delaunois B, Jeandet P, Clement C, Baillieul F, Dorey S, Cordelier S: **Uncovering plant-pathogen crosstalk through apoplastic proteomic studies.** *Frontiers in plant science* 2014, **5:**249.

27. Rastogi S, Rost B: **LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana.** *Nucleic acids research* 2011, **39:**D230-234.

28. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic acids research* 2003, **31:**365-370.

29. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.

30. Li S, Ehrhardt DW, Rhee SY: **Systematic analysis of Arabidopsis organelles and a protein localization database for facilitating fluorescent tagging of full-length Arabidopsis proteins.** *Plant physiology* 2006, **141:**527-539.

31. Foster LJ, de Hoog CL, Zhang Y, Xie X, Mootha VK, Mann M: **A mammalian organelle map by protein correlation profiling.** *Cell* 2006, **125:**187-199.

32. Mika S, Rost B: **UniqueProt: Creating representative protein sequence sets.** *Nucleic acids research* 2003, **31:**3789-3791.

33. Rost B: **Twilight zone of protein sequence alignments.** *Protein engineering* 1999, **12:**85-94.

34. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9:**56-68.

35. Bauer DC, Willadsen K, Buske FA, Le Cao KA, Bailey TL, Dellaire G, Boden M: **Sorting the nuclear proteome.** *Bioinformatics* 2011, **27:**i7-14.

36. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al: **Human Protein Reference Database--2009 update.** *Nucleic acids research* 2009, **37:**D767-772.

37. Wright PE, Dyson HJ: **Intrinsically disordered proteins in cellular signalling and regulation.** *Nature reviews Molecular cell biology* 2015, **16:**18-29.

38. Guy AJ, Irani V, MacRaild CA, Anders RF, Norton RS, Beeson JG, Richards JS, Ramsland PA: **Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions.** *PloS one* 2015, **10:**e0141729.

39. Vicedo E, Gasik Z, Dong YA, Goldberg T, Rost B: **Protein disorder reduced in Saccharomyces cerevisiae to survive heat shock.** *F1000Research* 2015, **4:**1222.

40. Nair R, Carter P, Rost B: **NLSdb: database of nuclear localization signals.** *Nucleic acids research* 2003, **31:**397-399.

41. Cokol M, Nair R, Rost B: **Finding nuclear localization signals.** *EMBO reports* 2000, **1:**411-415.

42. Wang Y, Zhang Q, Sun MA, Guo D: **High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles.** *Bioinformatics* 2011, **27:**777-784.

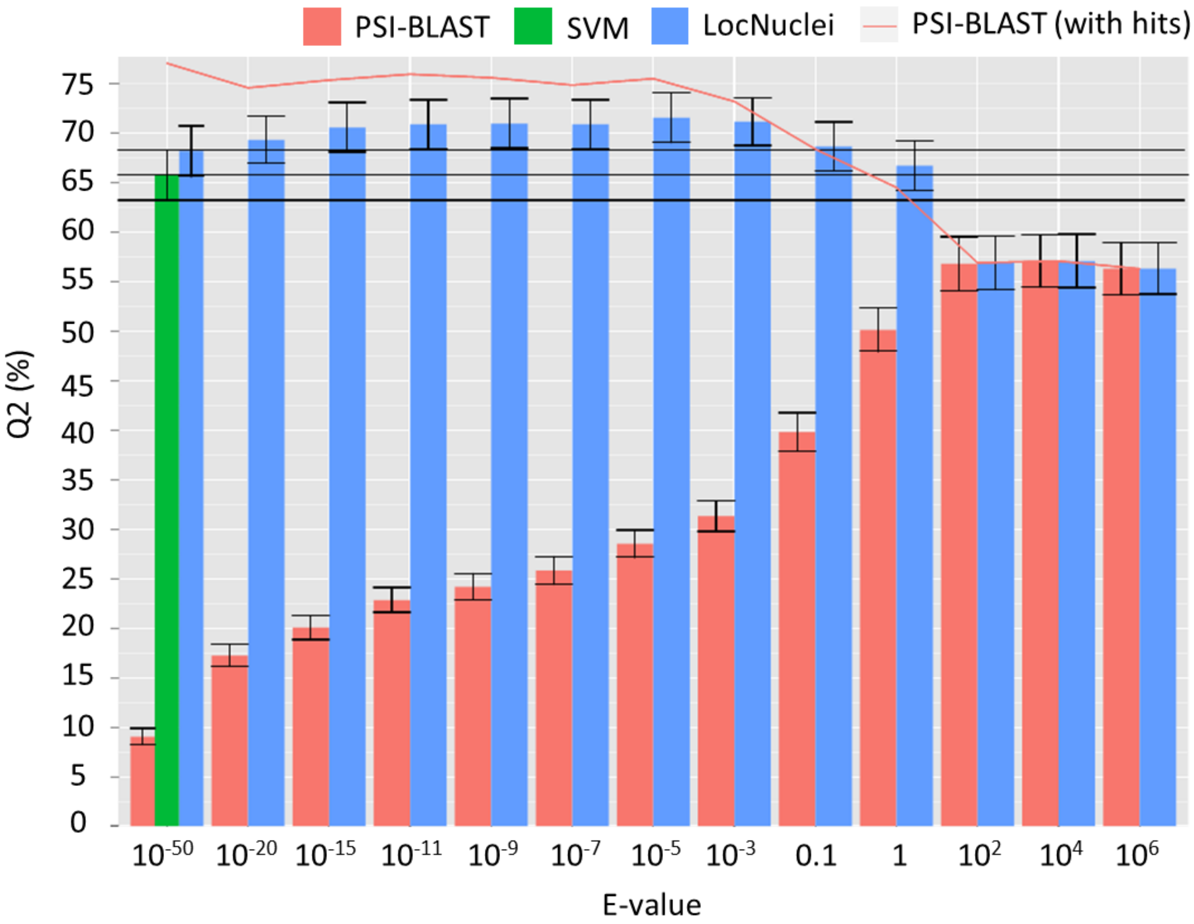
43. Wang Y, Sun M, Bao H, White AP: **T3_MM: a Markov model effectively classifies bacterial type III secretion signals.** *PloS one* 2013, **8:**e58173.

44. Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T: **Sequence-based prediction of type III secreted proteins.** *PLoS pathogens* 2009, **5:**e1000376.

9 Appendix

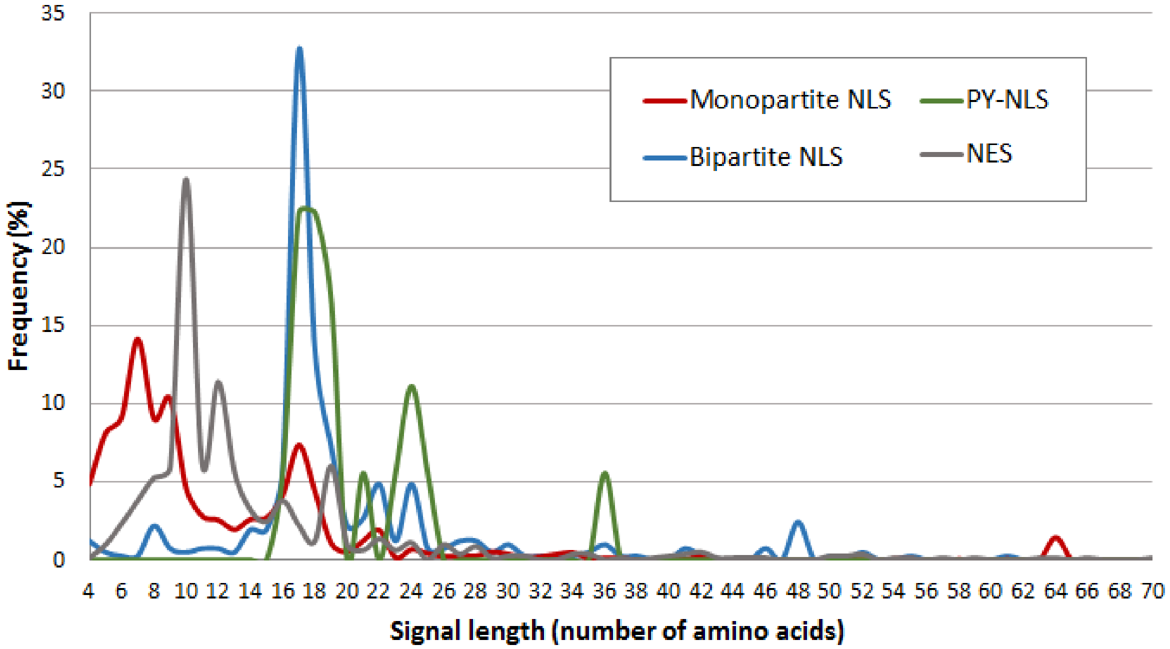
9.1 Supplementary Figures

Figure S1: E-value thresholds for the homology-based component of LocNuclei (prediction of nuclear travelers in two classes)



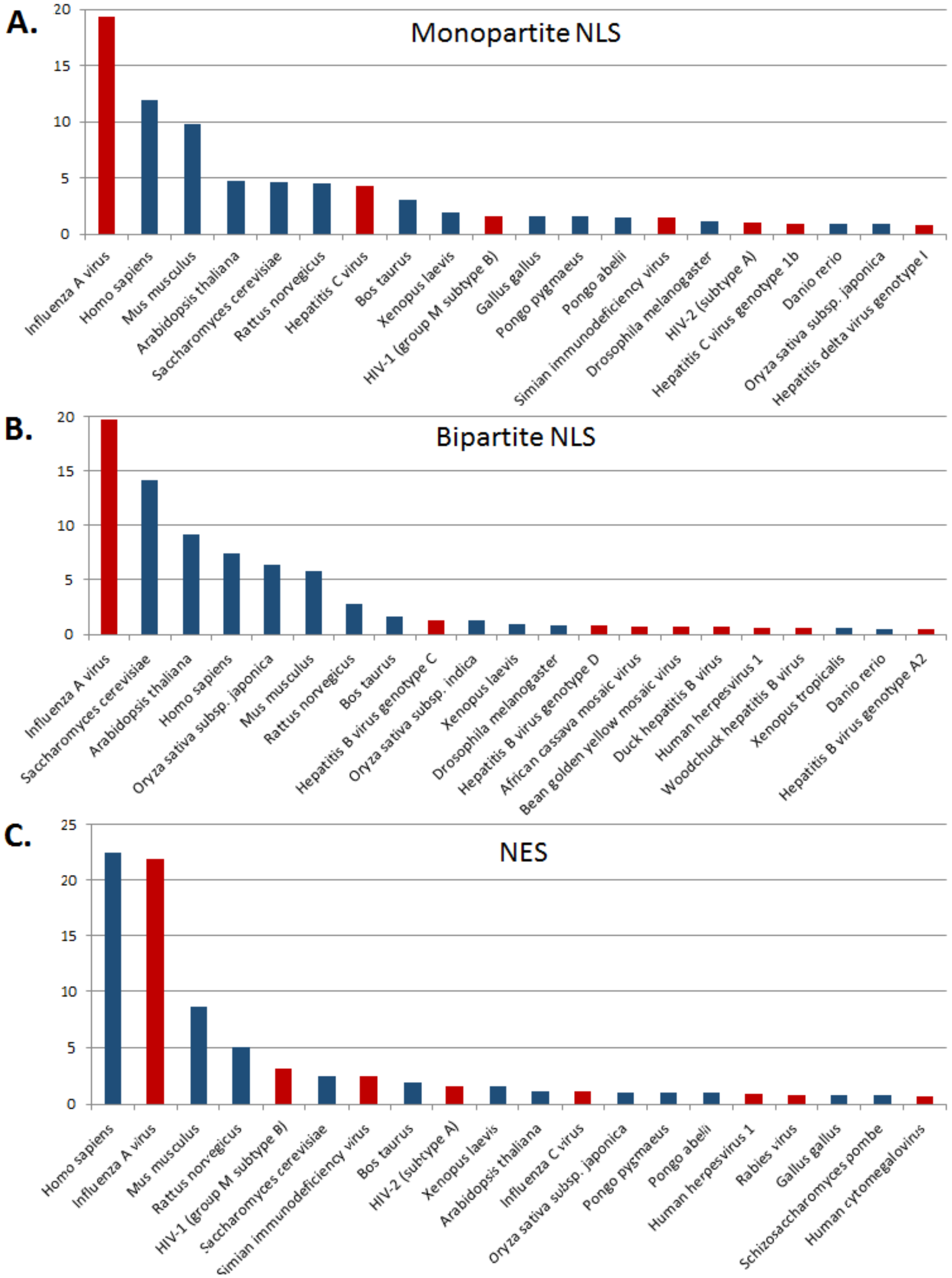
The figure shows the accuracy Q2 (Chapter 2.2) in classifying proteins in two classes (nuclear proteins exclusively localized to the nucleus and nuclear proteins localized also to other sub-cellular compartments) by LocNuclei and its components. The homology-based inference using PSI-BLAST from the set of experimentally annotated 12,055 nuclear proteins performs best (Q2 = 78%) at the stringent E-value ≤ 10^{-50}. However, when evaluated on the entire test set (*i.e.* also on proteins for which PSI-BLAST homology is not available), the performance drops significantly to Q2 = 9%. The performance of the SVM on the same set, however, reaches Q2 = 66% (the performance is marked by black lines). To determine, at which E-value threshold to use PSI-BLAST and at which the SVM, we needed to consider the performance of the final method LocNuclei at the same threshold. We found LocNuclei to reach highest Q2 = 72 ± 2% at E-value ≤ 10^{-5}.

Figure S2: Length distribution of all known nuclear signals



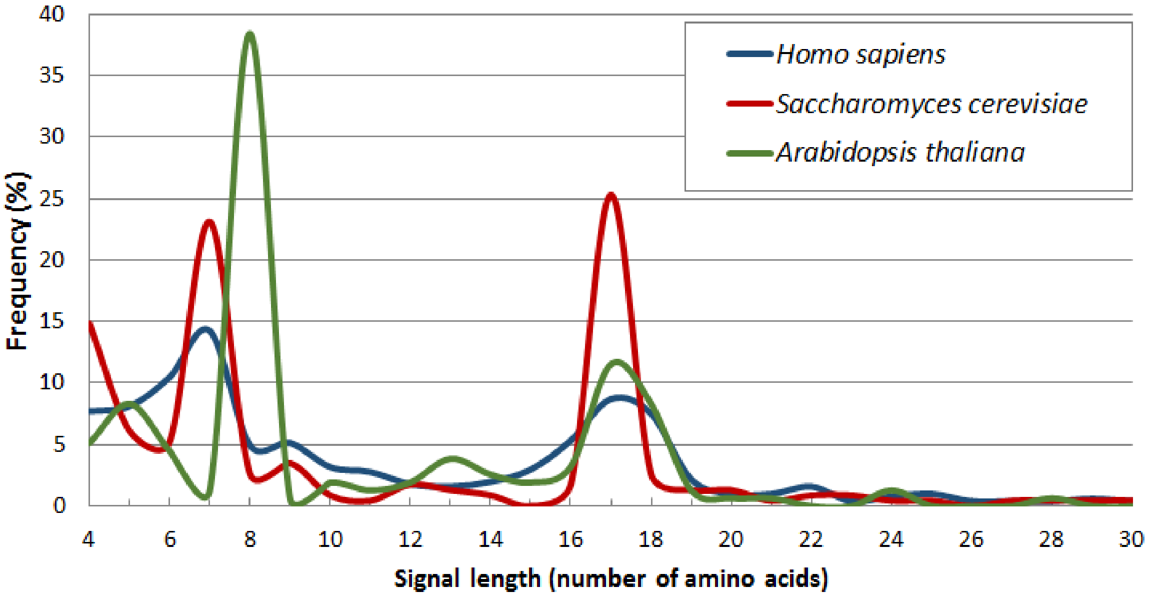
The figure shows the amino acid sequence length distribution of 1,960 monopartite NLS, 413 bipartite NLS, 18 PY-NLS and 817 NES in our trusted set (Chapter 5.2). The frequencies for each signal type sum up to 100%. Typical lengths for each signal type are represented by peaks: 4-10 and 15-19 amino acids for monopartite NLS (the latter is probably due to bipartite signals erroneously annotated as monopartite); 16-19 amino acids for bipartite NLS; 15-20 and 22-26 for PY-NLS; and 9-13 amino acids for NES.

Figure S3: Top 20 most frequent species annotations for nuclear signals



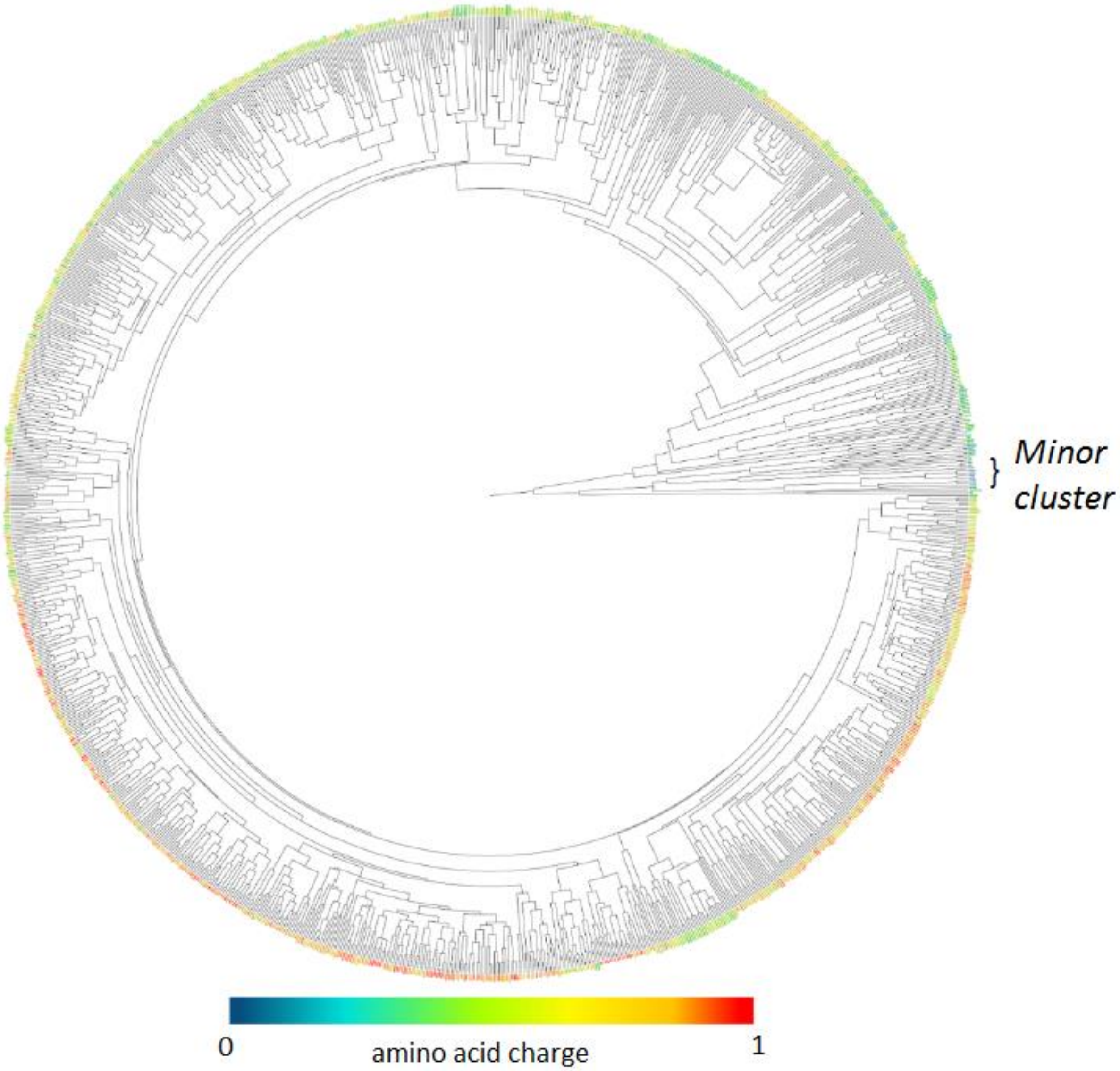
The figure shows top twenty of the most frequent organism species annotations for **(A)** monopartite NLS, **(B)** bipartite NLS and **(C)** NES. PY-NLS were annotated in human and Baker's yeast only and are not shown here. For each signal type, most frequent annotations were made in human and other eukaryotic model organisms, as well as in *Influenza A virus*. Virus species are colored red.

Figure S4: Length distribution of nuclear signals in human, yeast and plant



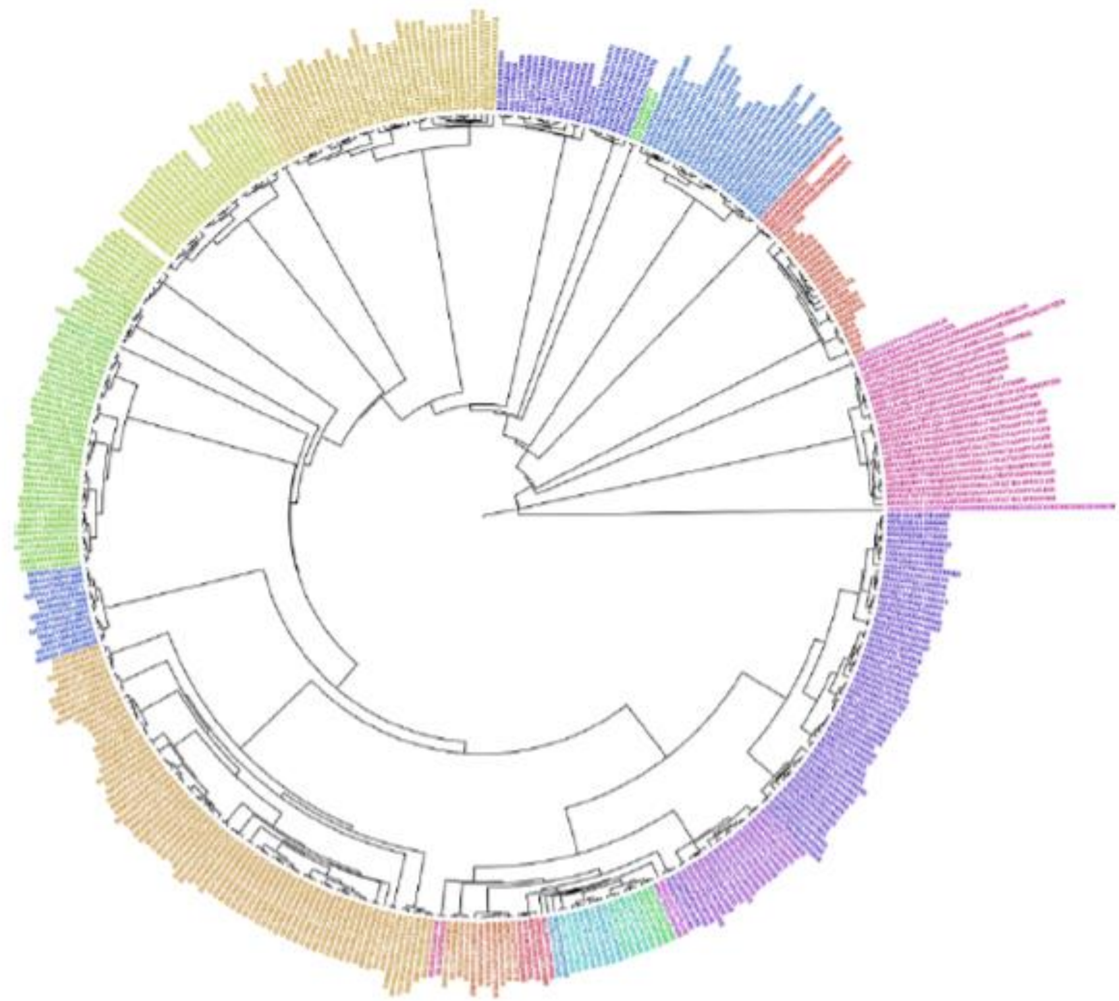
The curves show the distribution of unique nuclear signals (NLS and NES) annotated in human (*Homo sapiens*, 505 signals, blue line), yeast (*Saccharomyces cerevisiae*, 229 signals, red line) and plant (*Arabidopsis thaliana*, 156 signals, green line). The frequencies for each signal type sum up to 100%. Note we do not show results for sequences longer than 30 amino acids, as they constituted less than 1% of the data. For all organisms, the length of monopartite signals peaks in the range between 6 and 9 amino acids and of bipartite signals in the range between 16 and 19 amino acids. Monopartite signals appear to be most frequent in plant, while bipartite signals in yeast.

Figure S5: Phylogenetic tree representation of 1,960 monopartite NLS



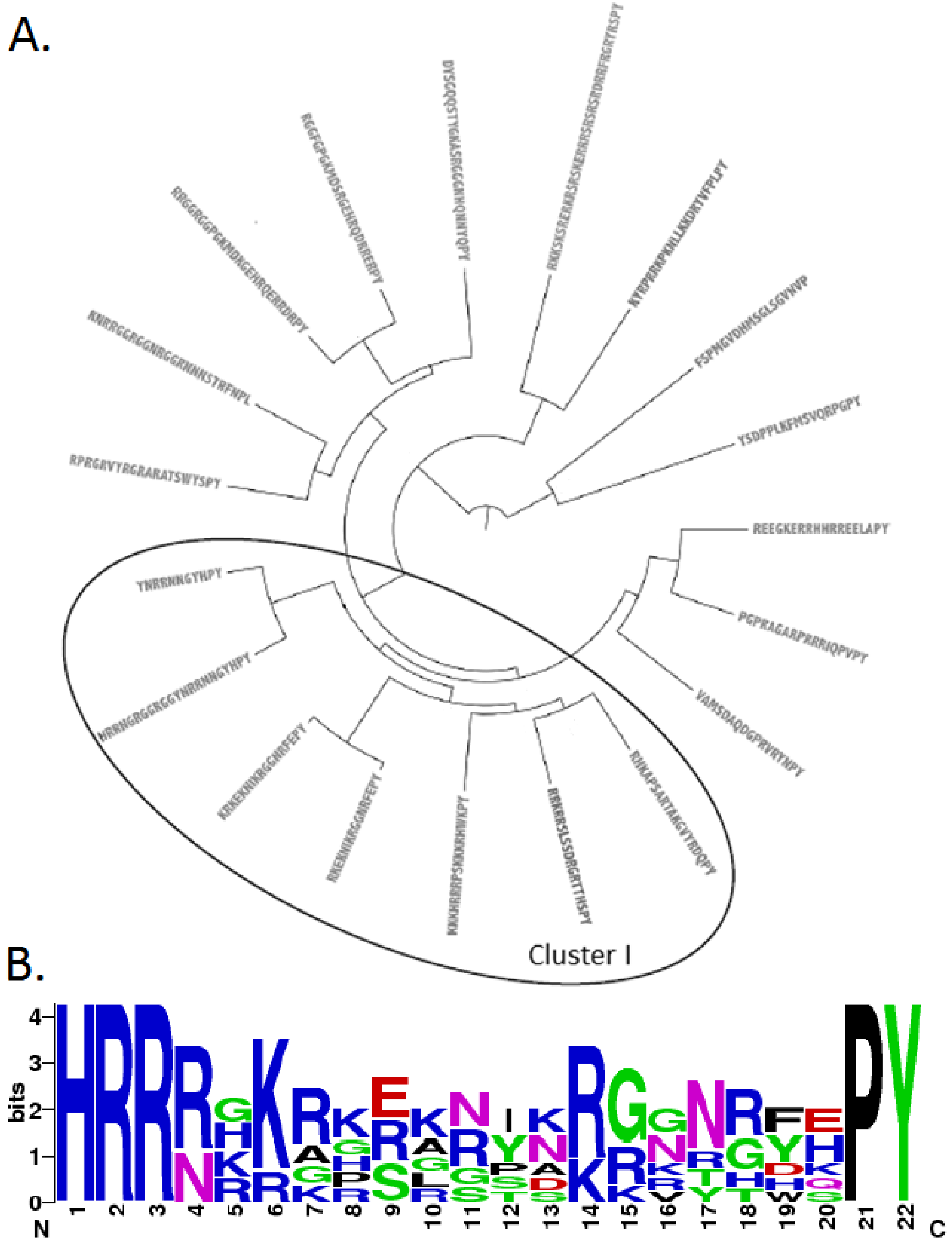
The sequences of 1,960 monopartite NLS from our trusted set (Chapter 5.2) were aligned against each other to construct an evolutionary distance matrix. This matrix was then used as input to the UPGMA clustering method [1] of the PHYLIP [2] package. The resulted phylogenetic tree separated all signals in two clusters: (i) “Minor” cluster of 13 sequences and (ii) Major cluster of all other sequences. The Major cluster was further sub-divided into 39 distinct clusters. Signal sequences in the tree are colored by the average charge of their amino acids. Only sequences of the “Minor” cluster appear to be negatively charged.

Figure S6: Phylogenetic tree representation of 413 bipartite NLS



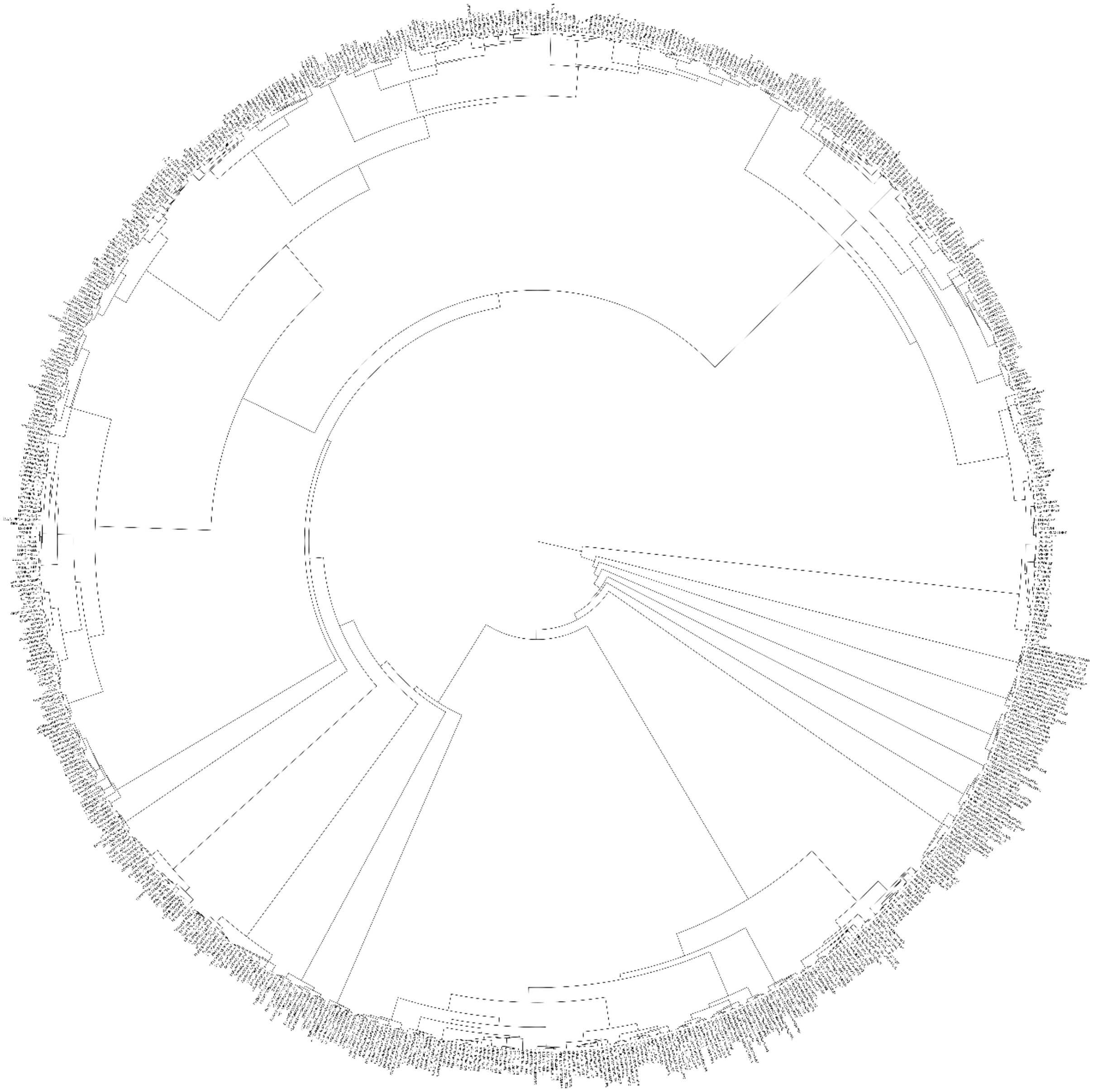
The phylogenetic tree for 413 sequences of bipartite NLS was constructed as in Figure S4. The bipartite signals formed 38 distinct sub-clusters. These are depicted by different colors.

Figure S7: Phylogenetic tree representation of 19 PY-NLS and the sequence logo of its largest sub-cluster



The data set of annotated PY-NLS comprised only 19 sequences, which were used to calculate the phylogenetic tree (**A**). The tree split the data in five clusters. The sequence logo representation (**B**) of the largest cluster (Cluster I) shows high conservation of amino acid residues at the flanking regions of the signal: basic residues in the N-terminal region and proline-tyrosine in the C-terminal region.

Figure S8: Phylogenetic tree representation of 817 NES



The phylogenetic tree for 817 sequences of NES was constructed as in Figure S4. The signals formed 27 distinct sub-clusters.

9.2 Supplementary Tables

Table S1: Normalization of sub-nuclear localization terms

Databases term	*Normalized term*
Cajal body, cajal bodies, gem	Cajal bodies
Chromatin, centromere, chromosome, heterochromatin, telomere, unsynapsed chromosome axes	Chromatin
Nuclear envelope, nuclear membrane, nucleus membrane	Nuclear envelope
Nuclear lamina, nuclear periphery, nucleus lamina	Nuclear lamina
Nuclear matrix, nucleus matrix	Nuclear matrix
Nuclear pore	Nuclear pore complex
Nuclear speckle	Nuclear speckles
Nucleolus, nucleolar	Nucleolus
Nucleoplasm	Nucleoplasm
Perinucleolar	Perinucleolar compartment
PML body, nuclear dots, PML-NBs, PML/ND10 bodies	PML bodies
Kinteochore	Kinetochore
Spindle apparatus, spindle microtubules, spindle midzone, spindle poles	Spindle apparatus

Databases HPRD [3], NMPdb [4], NOPdb [5], NPD [6], NSort/DB [7] and Swiss-Prot [8], annotate sub-nuclear proteins using synonyms for some terms. We extracted these terms and normalized them to 13 sub-nuclear localization classes. The normalization was done case-insensitive; terms of the same class are separated by comma.

Table S2: Composition of the sub-nuclear development set for LocNuclei

Sub-nuclear compartment (number of proteins per compartment)	*Chromatin (697)*	*Nucleolus (653)*	*Nuclear speckle (292)*	*PML body (95)*	*Nuclear lamina (80)*	*Nuclear matrix (74)*	*Nuclear envelope (72)*	*Cajal body (42)*	*Nuclear pore complex (35)*	*Nucleoplasm (29)*	*Kinetochore (25)*	*Spindle apparatus (14)*	*Perinucleolar (13)*
Chromatin (697)	**584**												
Nucleolus (653)	68	**483**											
Nuclear speckle (292)	22	79	**176**										
PML body (95)	23	18	9	**49**									
Nuclear lamina (80)	5	8	2	3	**51**								
Nuclear matrix (74)	4	3	3	2	1	**63**							
Nuclear envelope (72)	2	0	0	1	3	1	**63**						
Cajal body (42)	3	15	14	4	2	0	0	**15**					
Nuclear pore complex (35)	5	6	3	2	17	0	0	2	**12**				
Nucleoplasm (29)	3	8	3	1	0	2	2	0	0	**13**			
Kinetochore (25)	3	4	0	0	1	0	0	0	2	1	**15**		
Spindle apparatus (14)	2	1	0	1	4	1	1	0	1	0	3	**6**	
Perinucleolar (13)	3	4	6	2	1	0	0	1	0	0	3	0	**2**

The table displays numbers of sequence-unique proteins (HVAL [9, 10] ≤ 20) across 13 sub-nuclear localization classes in the development set of LocNuclei (Chapter 4). We only used proteins with experimental annotations extracted from HPRD [3], NMPdb [4], NOPdb [5], NPD [6], NSort/DB [7] and Swiss-Prot [8]. The numbers of unique sequences per localization are given in parentheses. The numbers on the diagonal describe sequences with the annotation of one localization class (*e.g.* 584 sequences in our set were annotated to localize at the chromatin only). Other numbers are annotations of two sub-nuclear compartments. Note that some sequences had annotations of more than two compartments.

9.3 References

1. Michener CD, Sokal RR: **A quantitative approach to a problem of classification.** *Evolution* 1957, **11:**490–499.
2. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
3. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al: **Human Protein Reference Database--2009 update.** *Nucleic acids research* 2009, **37:**D767-772.
4. Mika S, Rost B: **NMPdb: Database of Nuclear Matrix Proteins.** *Nucleic acids research* 2005, **33:**D160-163.
5. Leung AK, Trinkle-Mulcahy L, Lam YW, Andersen JS, Mann M, Lamond AI: **NOPdb: Nucleolar Proteome Database.** *Nucleic acids research* 2006, **34:**D218-220.
6. Dellaire G, Farrall R, Bickmore WA: **The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome.** *Nucleic acids research* 2003, **31:**328-330.
7. Willadsen K, Mohamad N, Boden M: **NSort/DB: an intranuclear compartment protein database.** *Genomics, proteomics & bioinformatics* 2012, **10:**226-229.
8. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic acids research* 2000, **28:**45-48.
9. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9:**56-68.
10. Rost B: **Twilight zone of protein sequence alignments.** *Protein engineering* 1999, **12:**85-94.