Lehrstuhl für Genomorientierte Bioinformatik,
Institut für Bioinformatik, GSF - Forschungszentrum für Umwelt und Gesundheit

New Approaches in Context-based Gene Function Prediction

Grigory Kolesov

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer.nat.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.rer. nat.habil. D. Langosch
Prüfer der Dissertation:

1. Univ.-Prof. Dr. rer.nat. H.-W. Mewes
2. Univ.-Prof. Dr. med., Dr. med. habil. Th. Dandekar, Bayerishe Julius-Maximilians-Universität Würzburg

Die Dissertation wurde am 18.11.2004 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 11.03.2005 angenommen.

New approaches in context-based gene function prediction

Summary

In this work we analyze genomic neighborhood of a gene as a source of functional information. Using concept of Similarity-Neighborhood graph we show presence of non-trivial relations between genomic neighbors in the context of several genomes. We also show that while for prokaryotes the existence of such relations can be deduced from the operonic organization of prokaryotic genomes, these relations can also be detected in the eukaryotic genomes where such organization is not common (with notable exception of *Caenorhabditis elegans* and possibly other members of *Nematodae*). We demonstrate applicability of our method for uncovering gene's function and studying properties of genomes.

We also demonstrate the techniques and tools developed for the analysis of genomic data. PEDANT genome system has been developed by our group and served as a main foundation for development of Similarity-Neighborhood approach, with the latter coming into life as a separate gene-function prediction tool - SNAPper web server. We discuss the techniques hiding behind our all-against-all protein alignment database, which has been developed as part of PEDANT genome analysis system. Such database is the requirement for nearly any cross-genome comparison approach, as it provides the basis for delineating of orthologous and paralogous groups of genes. One of such approaches - phylogenetic profiling, has been implemented by us in collaboration with Philip Wong and Walid Houry of University of Toronto, as highly flexible Web-based tool called PWP. The *Jaba* visualisation tool which we developed for manual analysis of genomes and multiple gene predictions and which has been extensively used in several genome projects, including large eukaryotic genome projects such as *Arabidopsis thaliana* and *Neurospora crassa* is also presented.

Briefly, Remm *et al.* define in-paralogous genes as genes which undergone duplication event after species' split (and therefore are essentially orthologous to their counterparts in another specie) in contrast to out-paralogs whose duplication precedes the speciation.

Detection of orthologs as implemented in INPARANOID starts with calculation of all-against-all alignments. This step usually performed by using BLAST (Altschul *et al.*,1998), because of its speed and because it is well-established software. Pairwise similarity scores for a pair of genomes (protein sets) A and B are calculated (A vs B, B vs A) as well as self-scores (A vs A, B vs B). When out-group protein set C is used, A vs C and B vs C are computed as well. Bit scores are made symmetrical by averaging reciprocal scores.

On the next step reciprocal best BLAST matches are found. Such pair of proteins from genome A and B is then considered as a central orthologous pair, around which additional orthologs (in-paralogs) are clustered. The basic assumption made by Remm *et al.* is that in-paralogs are more similar to the main ortholog, than to any sequence from other species.

On the last stage of algorithm the overlaps are resolved using different rules depending on the type and extant of the overlap; confidence values are computed using bootstrap.

The outline of INPARANOID algorithm is depicted on Fig. 1.1. The method of clustering in-paralogous into orthologous groups is visualized on Fig. 1.2.

Remm *et al.* compared the results of INPARANOID run to the results obtained from manually curated analysis of phylogenetic trees computed with nine different approaches. As it turns out, while INPARANOID failed to report less than 3% of orthologs found using these methods it reported considerable amount of additional orthologs, which may represent false positives or novel true orthologs. As it has been demonstrated by Remm et al, in many cases BLAST is able to detect more orthologs than phylogenetic approaches, due to its better sensitivity and different treatment of gaps.

Better performance of INPARANOID in comparison to COGs in terms of separation of in-paralogs and out-paralogs has been also shown. In large part this is due to the fact that COGs operate on the set of species larger than two, while INPARANOID is essentially two-lineage approach which allows it to precisely define the evolutionary point of orthology. At the same time that is also a drawback of the INPARANOID approach: for instance, in our case we would have to implement additional clustering techniques to cluster INPARANOID orthologs in COG-like groups. Preliminary studies conducted by us show that it is generally not very overwhelmingly difficult task though: INPARANOID pairwise ortholog groups have very low noise ratio which makes clustering them into multi-species groups relatively simple. In other words, simple approaches such as single-linkage clustering do not result in one huge cluster which encompasses everything.

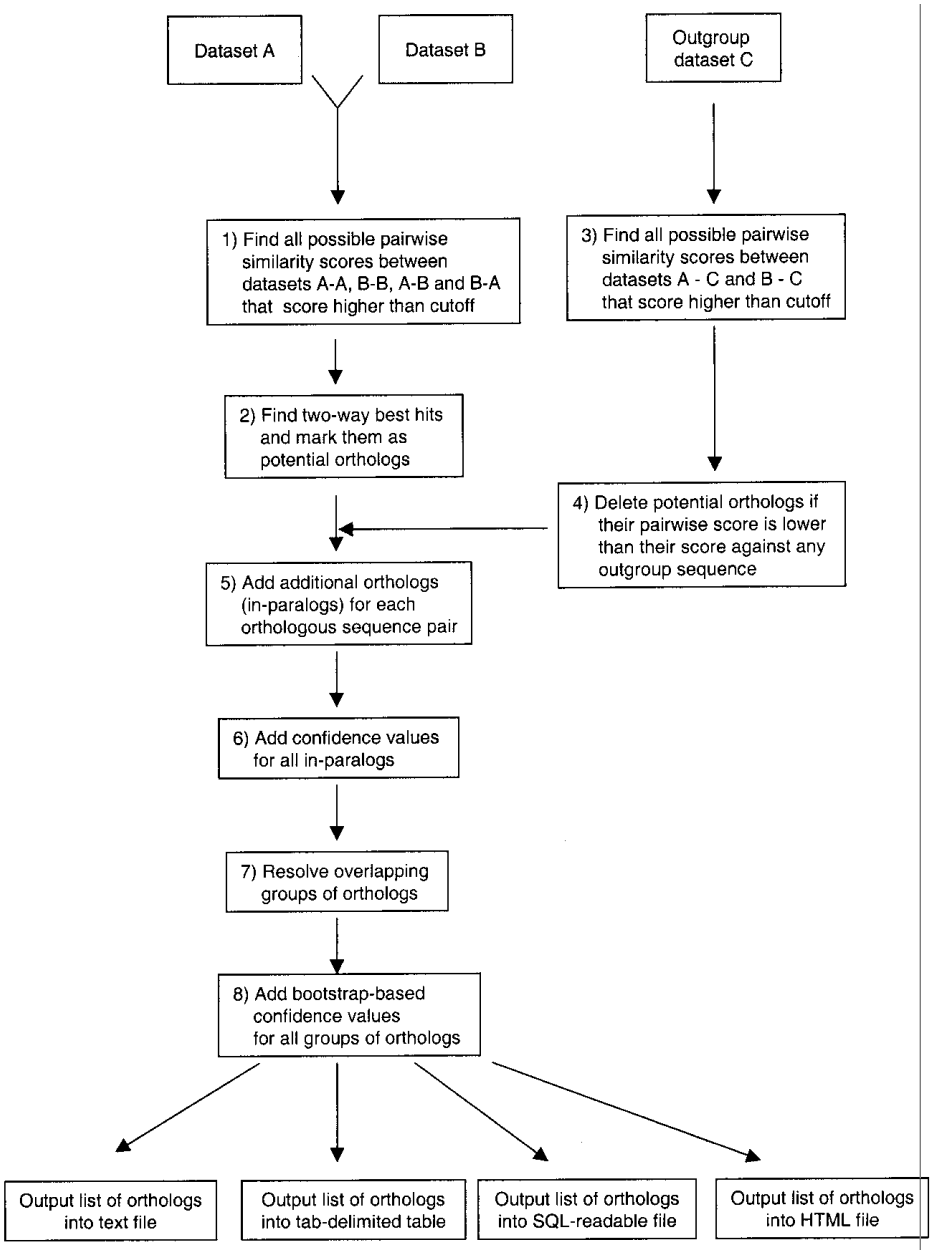


Figure 1.1. Overview of the INPARANOID algorithm. The program requires two fasta format sequence files A and B with protein sequences. All-versus-all BLAST search is run (1) and sequence pairs with mutually best hits are detected (2). Sequences from outgroup species are optionally used to detect cases of selective loss of orthologs. The A-B sequence pairs are eliminated if either sequence A or sequence B scores higher to outgroup sequence than they score to each other (3,4). Additional orthologs (in-paralogs) are clustered together with each remaining pair of potential orthologs as shown in Figure . Overlapping clus- ters are resolved by a set of rules. Finally, the bootstrapping technique is used to estimate the probability that a given pair of orthologs had mutual best score only by chance (8). The bootstrapping step is optional. (From Remm *et al*).

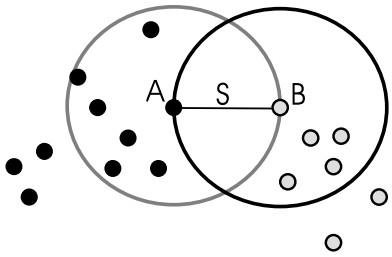


Figure 1.2. Clustering of additional orthologs (in-paralogs). Each circle represents a sequence from species A (black) or species B (grey). Main orthologs (pairs with mutually best hit) are denoted A1 and B1. Their similarity score is shown as S. The score should be thought of as reverse distance between A1 and B1, higher score corresponding to shorter distance. The main assumption for clustering of in-paralogs is that the main ortholog is more similar to in-paralogs from the same species than to any sequence from other species. On this graph it means that all in-paralogs with score S or better to the main ortholog are inside the circle with diameter S that is drawn around the main ortholog. Sequences outside the circle are classified as out-paralogs. In-paralogs from both species A and B are clustered independently (from Remm *et al.*).

1.1.2. COGs - Clusters of Orthologous Groups

COGs database (Tatusov et al. 2000) is an application of orthology concept to sets of multiple genomes. Originally COGs were developed for prokaryotic genomes but they have recently been extended to include large eukaryotic genomes such as *Caernorhabditis elegans* and *Drosophila melanogaster* (Tatusov *et al.* 2001).

COGs too are constructed using all-against-all gapped BLAST alignment. The underlying assumption of COGs approach is that any three proteins from the distant genomes that are more similar to each other than they are to any other proteins from the same genomes belong to an orthologous family. Further such minimal triangular clusters are extended by joining triangles sharing one edge. Thus, COGs unlike INPARANOID require at least three-species similarity relationships.

COG algorithm includes the following steps:

1 Perform the all-against-all protein sequence alignment.

2 Detect and collapse obvious in-paralogs, proteins from the same genome that are more similar to each other than to any proteins from other species. This step is similar to INPARANOID.

3 Detect triangles of mutually consistent, genome-specific best hits, taking into account the paralogous groups detected at step 2.

4 Merge triangles with common side to form COGs.

5 Manual analysis of each COG to remove false-positives.

6 Manual analysis of large COGs that include multiple members from all or several of the genomes using phylogenetic trees, cluster analysis and visual inspection of alignments. As result of this step, COG can be split to two or more consistent COGs.

COGs are arguably most used orthology database due to its effective and elegant clustering technique, manual curation and availability.

The drawback of using COGs is that it is a fixed database which includes limited number of genomes rather than a dynamic algorithm such as INPARANOID. Although some ideas behind COGs can be used separately to construct orthologous families.

Another drawback of COGs which it shares with INPARANOID is that horizontal gene transfer is not taken into account, although it has been shown relatively widespread among prokaryotes (Omelchenko *et al.* 2003; Brochier *et al.* 2000) and in some cases it can probably have dramatic effect on construction of orthologous groups.

1.2. Phylogenetic profiles

Biochemical pathways are not rigid constructs it terms of their occurence in various organisms. Some of them can be fairly conserved, some can disappear in the course of evolution, some can get displaced by alternative routes. The presence of certain biochemical route is determined by the presence of certain enzymes; coordinated action of these enzymes builds up that route. Therefore if in an organism we observe the enzyme belonging to the particular reaction path we can assume that the other enzymes participating in the path are likely to be present in that organism as well. Even more extreme example of such correlated occurrence is when a protein participates in multi-subunit complex and all parts of the complex have to be present in the organism to render the complex functional.

Recently suggested method of phylogenetic profiling (Pellegrini *et al.* 1999; Marcotte *et al.* 1999) explores a related idea: if two genes are not present individually in any of the genomes, *i.e.* the presence of one gene in the genome always implicates the presence of another, these genes are functionally related. The hypothesis behind this statement is that functionally linked proteins evolve in a correlated fashion, and, therefore, they have homologs in the same subset of organisms.

As a formalization of this approach, a bit-string of size N is constructed for each protein, where N is a number of genomes in the set. The presence of protein's homolog in the ith genome is indicated by setting ith bit of the string to 1. The bit string is essentially a phylogenetic profile of a particular protein. Further, proteins are clustered according to the similarity of their bit strings (Figure 1.3). The prediction method is based upon the assumption that functions of the proteins within one cluster are likely to be similar. Thus by using this method one can assign function (broadly) to the uncharacterized proteins using known proteins in the cluster.

The relevance of the method has been demonstrated on the real data. For instance, let us

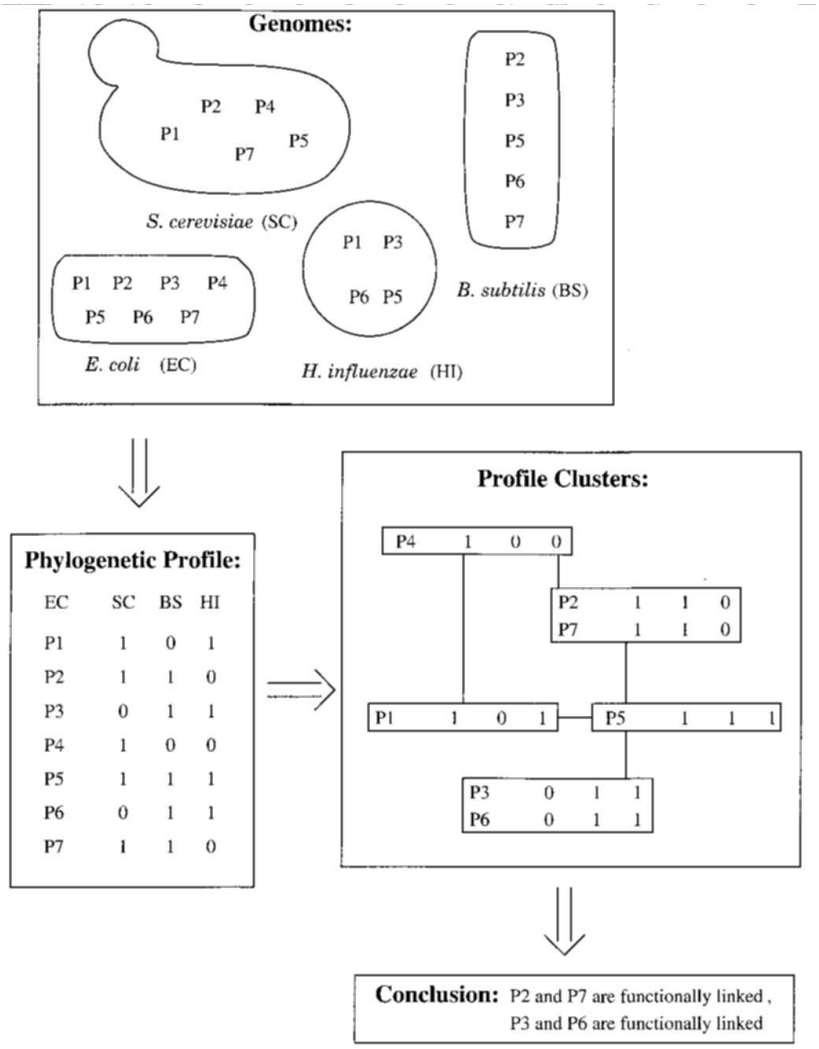


Figure 1.3. The method of analyzing protein phylogenetic profiles is illustrated schematically for the hypothetical case of four fully sequenced genomes (from *E. coli*, *Saccharomyces cerevisiae*, *Haemophilus influenzae*, and *Bacillus subtilis*) in which we focus on seven proteins (P1–P7). For each E. coli protein, we construct a profile, indicating which genomes code for homologs of the protein. We next cluster the profiles to determine which proteins share the same profiles. Proteins with identical (or similar) profiles are boxed to indicate that they are likely to be functionally linked. Boxes connected by lines have phylogenetic profiles that differ by one bit and are termed neighbors (from *Marcotte et al.*).

consider cluster of proteins with phylogenetic profiles no more than 1 bit different to the phylogenetic profile of ribosomal protein RL7(Figure 1.4(a)). This protein is present in nearly all eubacterias and yeast but not in *Archae*. Indeed, more than half of the proteins in the cluster are known to be associated with the ribosome.

The applicability of the method has also been shown on few other examples including histidin metabolism and flagella-related proteins (Figure 1.4(b,c)]). As it could be noted

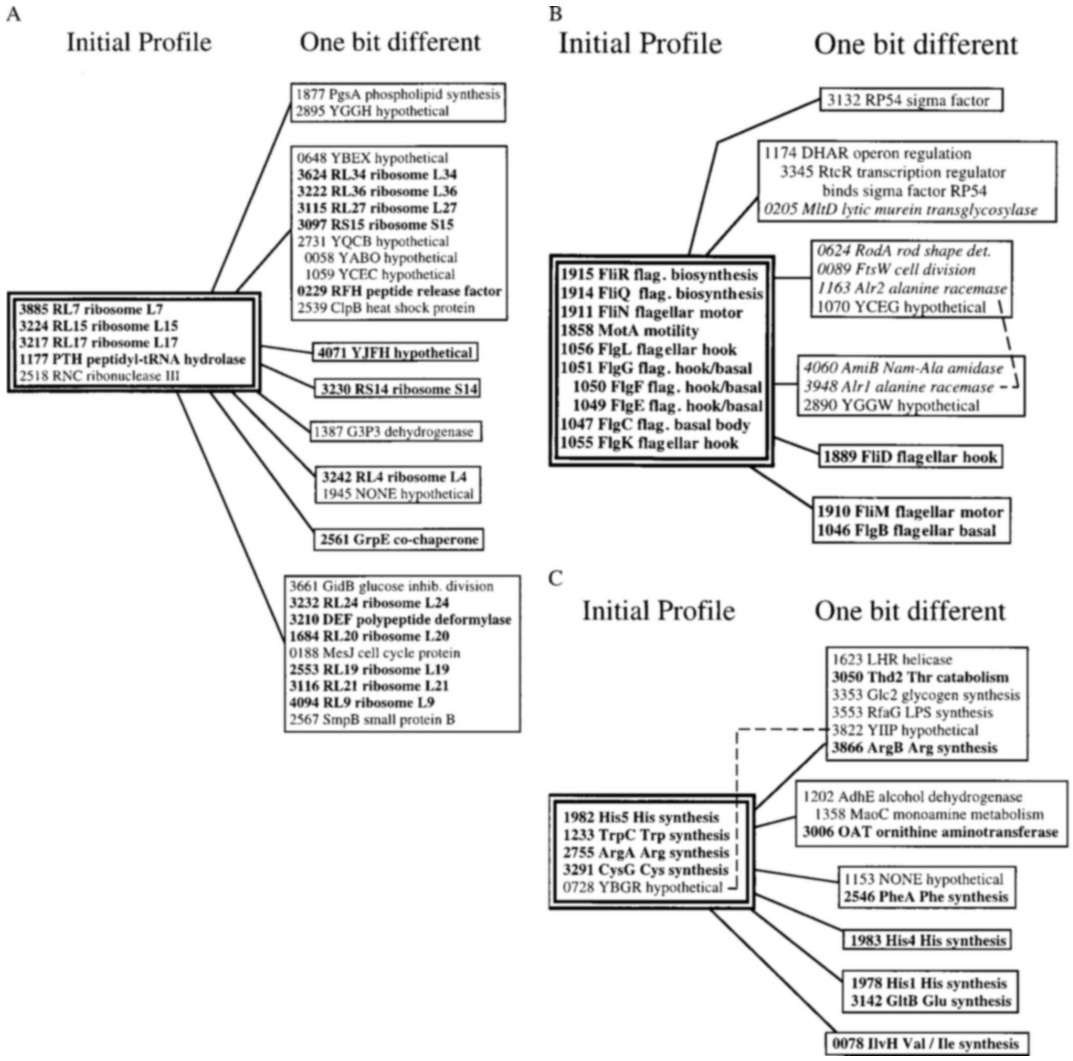


Figure 1.4. Proteins with phylogenetic profiles in the neighborhood of ribosomal protein RL7 (A), flagellar structural protein FlgL (B), and histidine biosynthetic protein His5 (C). In each case, we first found all proteins with profiles identical to our query proteins; the proteins we found are shown in the double boxes. We then found all the proteins with profiles that differed from our query proteins by one bit; these are shown in the single boxes. Proteins in bold participate in the same complex or pathway as the query protein, and proteins in italics participate in a different but related complex or pathway. Proteins with identical profiles are shown within the same box. Single lines between boxes represent a one-bit difference between the two profiles. All neighboring proteins whose profiles differ by one bit from the query protein are shown. Homologous proteins are connected by a dashed line or are indented. Each protein is labeled by a four-digit E. coli gene number, a SwissProt gene name, and a brief description. Note that proteins within a box or in boxes connected by a line have similar functions. Hypothetical proteins (i.e., those of unknown function) are prime candidates for functional and structural studies. Proteins in the double boxes in A, B, and C have 11, 6, and 10 ones, respectively, in their phylogenetic profiles, of a possible 16 for the 17 genomes presently sequenced (from *Marcotte et al.*).

all of the examples provided are focused on conserved biological mechanisms. This may hint to one of the drawbacks of the method: it is sensitive to the quality of recognition of orthologous proteins. Obviously, a protein wrongfully recognized as ortholog or unrecognized, but existing, ortholog, would affect phylogenetic profile adding 1 bit of difference to the targeted protein's profile. It can dramatically affect the quality of found clusters especially in the presence of phylogenetically distant genomes in the set, where homology relationships can not be as easily inferred. Therefore the value of the method for the characterization of less known, less studied in experiments (biologists tend to study core and often more conserved processes first), proteins is limited.

Another problem one has to be aware of when applying the method of phylogenetic profiles is that it is not always possible to separate process-specific genes from phylogenetic signal (taxon-specific genes). For instance, by clustering the phylogenetic profiles of the oxygenic photosynthesis-related proteins of *Cyanobacteria* species (the only prokaryotic organisms which are capable of this process) not only the proteins directly participating in the oxygenic photosynthesis could be extracted but other *Cyanobacteria*-specific proteins as well.

On the other hand, it can be useful in certain cases to restrict phylogenetic profiling to the certain taxon. For example, phylogenetic profiling predicts functional linkage between GroEL and GroES if only *Bacteria* species are present in the set. While *Archaea* species do have GroEL, they do not possess GroES homolog and therefore adding them to the genome set would hamper proper clustering of GroEL's and GroES' phylogenetic profiles.

An on-line resource for phylogenetic profiling has been developed recently (Wong *et al.* 2003) by our group (see Chapter 3).

More elaborate method which incorporates phylogenetic information in the analysis has been published recently (Vert 2002). In this paper author generalizes the approach of phylogenetic profiles and builds up powerful mathematical framework to analyze them. Instead of just defining bit difference-based measure of similarity between profiles author suggests to map profiles in higher-dimensional features space. This feature space is defined in such way that each point in the space corresponds to a pattern of inheritance during evolution. An example such pattern could be "this gene has been transmitted to proteobacteria and eubacteria but not to Gram-positive bacteria".

As it is impossible to know the exact content of ancestral genomes, the mapping of phylogenetic profiles to the feature space is defined using a probabilistic (Bayesian) model of evolution giving weights to 'features' which correspond to plausible patterns of inheritance for a particular profile.

As there exists immense number of possible patterns of evolution the dimension of the feature space is very large; consequently, the explicit computation of the image of a profile could be infeasible. However, the method was provided to efficiently compute the inner product between the images of any two profiles in the feature space.

The function designed to map any two phylogenetic profiles to the inner products of their images in the high-dimensional feature space is called *tree kernel*. It belong to a larger class of functions, called *kernels*, defined as the inner product of two objects mapped to any vector space. Once a kernel and a corresponding feature space are chosen it is possible to define the Euclidian distance between any two images of the phylogenetic profiles in the feature space. In this case, two profiles are near in the feature space if they are likely to share common pattern of inheritance during evolution, which is an appealing property.

However the application of kernel function is not limited to computing of Euclidean distances. Whole set of new algorithms, known as *kernel methods*, can be applied once the kernel function is defined. Kernel methods work implicitly in the feature space using only kernel function and include such popular algorithms as Support Vector Machine (SVM) (Vapnik 1998), kernel principal component analysis (Schölkopf *et al.* 317), kernel clustering (Ben-Hur *et al.* 2001) and Fisher discriminants (Mika *et al.* 1999).

Author applied SVM to infer the function of a gene from its phylogenetic profiles. SVMs are a class of supervised learning algorithms. Given a kernel *K(.,.)* and a set of a training examples (phylogenetic profiles) labeled as positive or negative, SVMs learn a linear decision boundary in the feature space defined by the kernel in order to discriminate between positive and negative examples. Any new unlabeled examples is then predicated to be positive or negative depending on the position of the image of phylogenetic profile in respect to a linear boundary obtained by SVM.

Author then tested a trained SVM using naive kernel (bit-distance based) and tree kernel on a set of proteins from *S. cerevisiae* genome. As it has been shown prediction power achieved by tree kernel surpasses by more than two folds that of bit distance-based approach at low values of false positives.

1.3. Patterns of domain fusion

Fused gene constructs are often used in experiments for biochemical analysis and protein-purification technology (Buxlow 1990; Wales and Wild 1991). These experiments are strikingly similar to what is being observed happening in nature: as it turns out, gene fusion is not at all rare or exotic event in the course of evolution (Enright *et al.* 1999). The classical example of such event is the fusion of α- and β- chains of tryptophan synthetase from bacteria to fungi (Burns *et al.* 1990), depicted on the Figure 1.5.

It is intuitively appealing to assume that proteins involved in the fusion events interact in the living cell. Indeed, the precision of such prediction of ~75% has been estimated. (Enright *et al.* 1999) The false positives and false negatives are believed largely due to the errors in the determination of protein-domain homology relationships. Thus precision of the method can be further increased by raising similarity cut-off values. The method's maximum coverage of 95% has also been estimated.

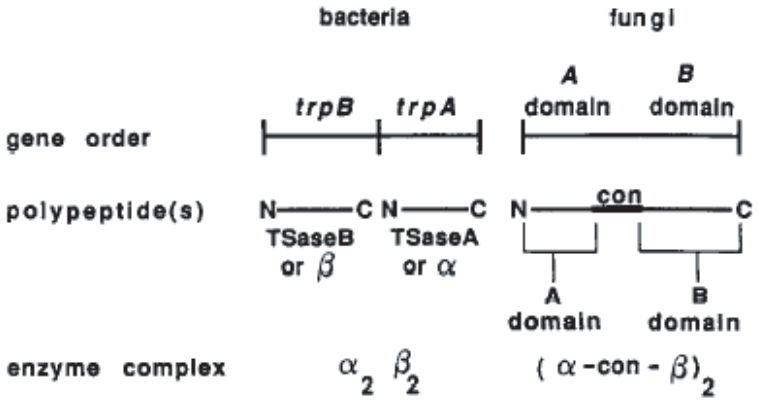


Figure 1.5. Schematic representation of tryptophan synthetase genes and polypeptides in bacteria and fungi. N and C refer to the amino- and carboxyl-terminal residues, respectively. Con refers to the connecting region in fungi that joins the TSase A and B domains (from *Burns et al.*).

The examples of gene fusion events are provided in the Table 1.1. All of the involved proteins are known interaction partners.

It has also been shown that non-neighbors (genes located far apart on bacterial chromosome) are as often involved in the fusion events as neighboring genes. This is an extremely important observation and we will return to it later.

Table 1 The 64 fusion events in the genomes of *E. coli*, *H. influenzae* and *M. jannaschii* detected on the basis of composite proteins in these three genomes and the genome of *S. cerevisiae*.

Case	Component	Component	Composite	EC	HI	MJ	SC	N
1	GalE	GalM	GAL10	▲▼	▲▼		●1	2
2	AccC	B0712-hypothetical	DUR1,2	▲▼	▲▼		●1	2
3	Hypothetical	Hypothetical	PYC2,PYC1			▲▼	●2	1
4	HisH	HisF	HIS7	▲▼	▲▼	▲▼	●1	3
5	HisI(E)	HisD	HIS4	▲▼	▲▼	▲▼	●1	3
6	RpoA′	RpoA″	RPO21,RPO31,RPA190			▲▼	●3	1
7	GltB	GltD	GLT1	▲▼			●1	1
8	AroB/AroA/AroK/AroD/AroE	Multiple fusion	ARO1	▲▼▲▼▲	▲▼▲□□	□▼□□▲	●1	3
9	Aconitase subunit	Aconitase subunit	LYS4			▲▼	●1	1
10	ArgA	ArgC	ARG5,6	▲▼			●1	1
11	LeuC	LeuD	LEU1	▲▼	▲▼	▲▼	●1	3
12	TrpA	TrpB	TRP5	▲▼	▲▼	▲▼	●1	3
13	PurD	PurM	ADE5,7	▲▼	▲▼	▲▼	●1	3
14	PurL	PurQ	ADE6	●1	●1	▲▼	●1	1
15	CarA/CarB/PyrB	Multiple fusion	URA2	▲▼▲		▲▼▲	●1	2
16	B1378	CysI	ECM17	▲▼			●1	1
17	TrpG	TrpC	TRP3	▲▼	▲▼	▲▼	●1	3
18	AgaG	AgaF	HyuA,HuyB			▲▼	●1	1
19	IlvG_1	IlvG_2	ILV2	▲▼	●1	●2	●1	1
20	GmpA	GmpB	GUA1	●1	●1	▲▼	●1	1
21	GyrB,ParE	GyrA,ParC	TOP2	▲▲▼▼	▲▲▼▼		●1	4
22	FolK	FolP	Folate biosynthesis (probable)	▲▼	▲▼▼		●1	3
23	PabA	PabB	ABZ1	▲▼	▲▼		●1	2
24	PurK	PurE	ADE2	▲▼	▲▼	▲▼	●1	3
25	RpoB″	RpoB′	RPB2,RET1,A135	●1	●1	▲▼	●3	1
26	ThiE	ThiM	THI6		▲▼		●1	1
27	ThiD	TenA	thi21,thi20,thi22		▲▼		●3	1
28	TktA	TktB	TKL1,TKL2	●2	●1	▲▼	●2	1
29	LysC	hom	ThrA,MetL	●2	●1	▲▼		1
30	ABC transporter	Hypothetical	B0879	●1		▲▲▼▼		4
31	Hypothetical	Putative methyltransferase	B0948	●1		▲▼▼		2
32	TrpG	TrpD	TrpD	●1	▲▼	▲▼		2
33	FumA	FumB	FumA	●1		▲▼		1
34	Hypothetical tkt	PheA	PheA	●1	●1	▲▼		1
35	FprA	Rubredoxin	B2710	●1		▲▲▲▼▼		6
36	TrxM	Hypothetical	B0492	●1	▲▼			1
37	Hypothetical	Hypothetical	B1816,B2063	●2	▲▼▼▼			3
38	CpxR,YgiX	TyrR	AtoC,YfhA,GlnG,HydG	●4	▲▲▼			2
39	Hypothetical	Multiple fusion	B2324	●1	▲▼▲			1
40	Hypothetical	Hypothetical	B2474	●1	▲▼			1
41	Hypothetical	Hypothetical	SufI	●1	▲▼			1
42	HemX	Hypothetical	HemX	●1	▲▼			1
43	TrpC	TrpF	TrpC		●1	▲▼		1
44	CitX	CitG	CitG	▲▼	●1			1
45	SbmA	Hypothetical	ABC transporter/ATP-binding	▲▼▼▼▼▼▼▼	●1			7
46	B3777	B3776	Hypothetical	▲▼	●1			1
47	B2612	YfjD	Hypothetical	▲▼	●1			1
48	YgfQ	YgfR	Hypothetical	▲▼	●1	●1		1
49	YabK	B0263	Hypothetical	▲▼	●1			1
50	YhaQ	YhaP	SdaA	▲▼	●1			1
51	YbfH	YbfG	Hypothetical	▲▼	●1			1
52	PurF	YhfN	GlmS	▲▼	●1			1
53	FrwB,FrwD	FrwC,B2386	FruA	▲▲▼▼	●1			4
54	UgpC	YtfS	RbsA,MglA	▲▼	●2			1
55	B1515,B1899	AraH	RbsC,MglC	▲▲▼	●2			2
56	NrfF	NrfG	NrfF	▲▼	●1			1
57	MsrA	B1778	MsrA	▲▼	●1			1
58	SbmA	Hypothetical	ABC transporter/ATP-binding	▲▼▼▼▼▼▼▼	●1			7
59	YhgK	YhgJ	Probable RNA cyclase	▲▼		●1		1
60	FrdB	GlpC	Iron-sulfur-binding reductase	▲▼		●1		1
61	RffH,RfbA,GalF,GalU	B0359	Glucose-1-P thymidyltransferase	▲▲▲▲▼		●1		4
62	B3016	B3015	Hypothetical	▲▼		●1		1
63	LeuS	YgjH	MetS	▲▼		●1		1
64	TopB	YrdD	TopA	▲▼	▲▼	●1		2
								122

The component gene/protein names (or identifiers) and the composite (fusion) gene/protein names (or identifiers) are listed. Columns EC, HI, MJ and SC correspond to *E. coli*, *H. influenzae*, *M. jannaschii* and *S. cerevisiae*, respectively; N lists the maximum number of possible pairwise interactions taking into account paralogy in the query genomes (multiple-component cases are counted as a single interaction). Symbols represent a corresponding component or composite genes/proteins: triangle pairs, ▲▼, a pair of component proteins in the query genome predicted to interact based on their similarity to a composite protein in the reference genome; alternating triangles, ▲▼▲/▲▼▲▼▲, multiple-component genes/proteins (cases 8, 15 and 39); open squares, □, absence of a component from a multiple-fusion event (case 8); consecutive triangles, ▲▲…/▼▼…, the exact number of detected paralogous component genes/proteins in the query genome; filled circles, composite genes/proteins, the number represents the number of paralogous composite genes/proteins in the reference genome. The sort order follows the three species against the composite-protein sequence identifiers for the yeast genome, and then the other three species in succession. Genes are named where possible; where none is available, the sequence identifier is used instead. All fusions were confirmed by reverse BLAST searches using the composite protein as query, which identified all the component proteins. Note that functional annotation is not necessary but frequently useful in resolving paralogous cases (for example, case 21). Predictions imply functional associations and not necessarily direct molecular interactions. For gene/protein identifiers and references for the known cases, see Supplementary Information.

1.4. Functional coupling of collinear gene pairs

Genes in prokaryotic chromosome are often appear grouped in the structures known as operons (Jacob *et al.* 1960) genome: it can contain several adjacent genes sharing

regulatory sites at 5' region of operon, transcription start and stop[5](see Figure 1.6). The presence of several downstream alternative transcription terminators as well as translation attenuators has been shown in some cases.

All genes in operon are transcribed by RNA-polymerase into single polycistronic mRNA, which is later translated by ribosome to produce individual polypeptides. The most prominent examples of operons are *E. coli*'s *Lac*-operon (Figure 1.6) (Jacob *et al.* 1960) , tryptophan operon and ribosomal operons.

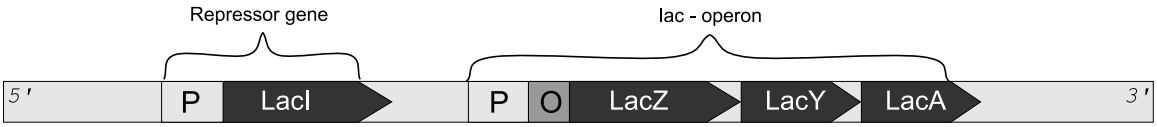


Figure 1.6. Schematic representation of *lac*-operon. *LacZ*, *LacY* and *LacA* are adjacent structural genes coding for proteins contributing to lactose transport and metabolism: β-galactosidase, galactoside permease and transacetylase, respectively. Structural genes are prefixed by promoter (P) and operator (O) sites. The product of *LacI* gene is a transcription repressor binding to the operator site of *lac*-operon and effectively blocking the transcription. In the presence of lactose lactose-repressor complex is unable to bind to the operator thus enabling the transcription.

Several alternative hypothesis explaining this organization have been proposed: evolutionary advantage of having a single point of regulation for genes functionally related to each other, e.g. participating in the same metabolic pathways; evolutionary pressure to group genes into functional modules to ease horizontal transfer of advantageous gene cliques (*selfish operon* model) (Lawrence and Roth 1996).

The latter hypothesis has gained large popularity among biologists and evolutionists despite of the lack of rigorous studies which would confirm or discard it and despite it's claim to be the only mechanism keeping functionally related genes in operon being in conflict with the gene shuffling observed in genomes.

Brochier *et al.* studied evolutionary fate of ribosomal protein RpS14 (Brochier *et al.* 2000). This protein belongs to the group of the most conserved proteins - it is necessary for assemblage of 30S ribosomal subunit and it is a part of peptide environment of the peptidyl transferase center, which is involved in the essential process of peptide elongation.

Considering the RpS14's crucial role in what could be designated as the heart of cell's activity - in translation and taking into account the large number of physical interactions this protein is involved in, it would seem improbable for this gene to be transfered from other species without it's ribosomal partners.

In contrary to this common-sense assumption Brochier *et al.* identified several cases of insertion of single foreign RpS14 gene into genomes' native ribosomal operon.

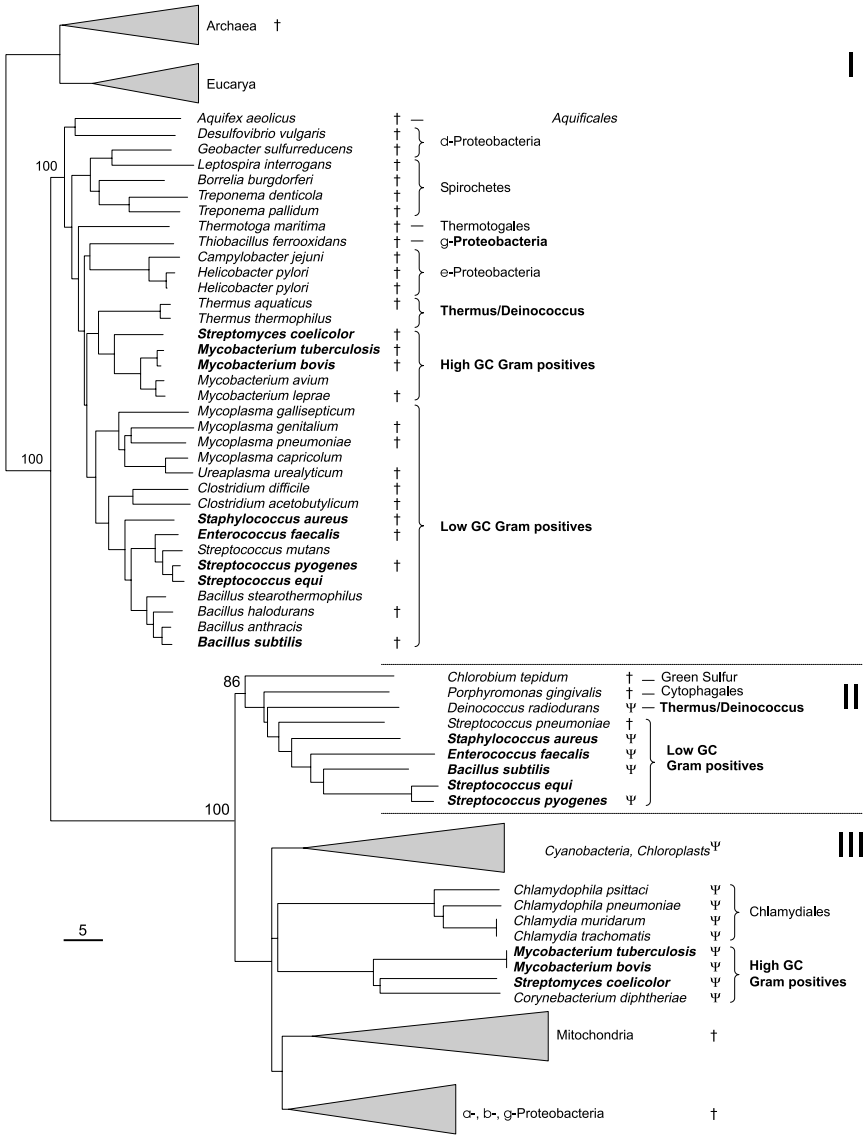


Figure 1.7. Bacterial *rps*14 sequences cluster within three main groups (I, II and II) in the phylogeny constructed using the neighbour-joining algorithm and rooted on the archaeal and eukaryotic sequences. Several monophyletic groups are displayed as solid triangles for clarity. Names of groups that are considered monophyletic on the basis of other phylogenetic markers but split in the RpS14 phylogeny are in bold. Species with two *rps*14 copies are in bold. The *rps*14 sequences mapping within *spc* canonical operon are indicated by †, those mapping within rearranged operons by Ψ. Bootstrap proportions estimated using 1000 replicatees are shown for the three main groups. The scale bar represents the number of substitutions per 100 sites for a unit branch length. The alignment, sequence accession number and the complete phylogenetic tree are available at htt://sorex.snv.jussieu.fr (from Brochier *et al*).

Analysis of phylogenetic tree of RpS14 depicted on Figure 1.7 showed that while some species occupy two distant branches in that tree, several others occupy branches distant

from their taxon which strongly suggests xenologous origin of their RpS14 gene.

One of the most remarkable large-scale horizontal gene transfer (HGT) of *rps14* gene is the acquisition of proteobacterial group III-type genes by Chlamydiales, Cyanobacteria and several high-GC Gram positive speices (Fig. 1.7). In all cases the acquired genes do not map to canonical for that gene *spc* operon but often to other ribosomal operons, which further strengths the hypothesis of *rps14*'s HGT origin (Figure 1.8). In one case (*Synechocystis sp.*) *rps14* seems to be isolated from other ribosomal genes but instead is located very closely to Arg-tRNA gene.

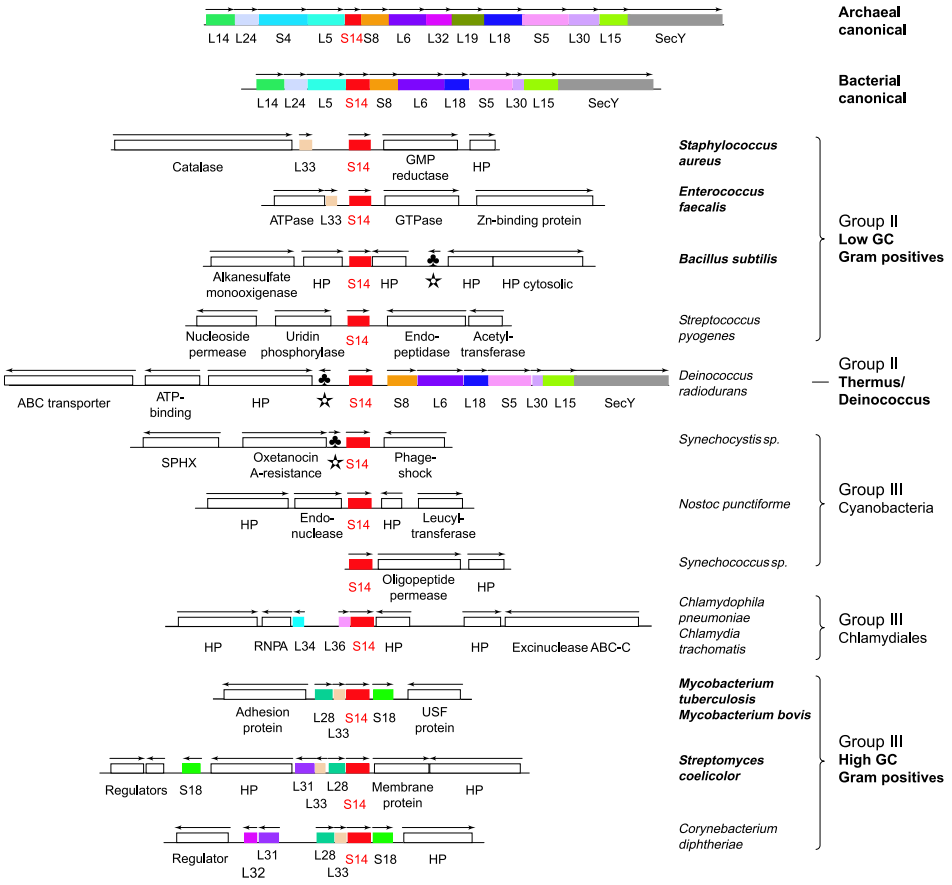


Figure 1.8. Archaeal and bacterial canonical *spc* operons, and genetic environments for the *rps*14 genes acquired by horizontal gene transfer (HGT) by several species. Boxes corresponding to genes for ribosomal proteins are in colour. tRNA genes are indicated by stars. Arrows indicate the sense of transcription. Species with two *rps*14 copies, one within a canonical operon and the other witin a rearranged operon, are in the old. Abbreviations: HP, hypothetical protein; RNPA, RNase P protein A. SPHX, periplasmic phosphate-binding protei; USF, putative carboxymethylenebutenolidase (from Brochier *et al*).

The HGT of *rps14* has been also observed in case of *D. radiodurans* , some Gram positive bacterias.

The authors conclude that it is difficult to explain multiple HGT events of single rps14 without predominant selective pressure favouring HGT. Authours propose that it could be that antibiotic resistance is conferred by the transferred sequences coming from resistant species (rps14 is known to be involved to antibiotic resistance).

This study discards the hypothesis that complexity of physical interactions in ribosome would prevent the transfer and integration of xenologous ribosomal protein into translation machinery. In addition, these results demonstrating the single gene being transferred from one specie to another and inserting into functionally close (ribosomal) operon but disrupting original operon for this gene (*spc*-operon) are questioning selfish operon model (SOM).

It has been shown in this work, an idea operon as just a unit of horizontal transfer does not match observations in the cases studied, which of course does not completely discard SOM. Also, the authors point out, only the cases that passed strict homology threshold have been considered.

In related study Daublin *et al.* challenge these and similar results which demonstrate the ubiquity of HGT in evolution of prokaryotic genomes (Daubin *et al.* 2003). The authors claim that such studies fall prey of certain problems and/or methodological artifacts: i) the complexity of phylogenetic analysis, namely overestimation of bootstrap support for gene trees and ii) wrong ortholog determination by using reciprocal best similarity hits as orthologous.

Authors go further and conduct analysis of gene trees computing the statistical support for them by applying Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999). Orthologous genes were selected by rather conservative approach: by including only those genes that have a single significant match per genome, thus minimizing the risk of including hidden paralogs descending from within genome duplication events. They analyze quartets of orthologous genes and compare the topology of resulting trees to topology of the tree derived from small subunit ribosomal RNAs (SSU rRNAs) of corresponding prokaryotic species. The species' quartets were subdivided into two groups in respect of the possibility of HGT: "Intraspecies" quartets which contained quartets of species and strains between which HGT is believed to be likely, and "Interspecies" quartets containing lineages for which HGT is unlikely due to the environmental (different ecological niche) or other reason.

The resulting frequencies of different quartet topologies are shown on Figure 1.9. The relative appearance of HGT topologies is very low even for intraspecies quartets. Authors examined the group of HGT topologies to discover that frequencies of such topologies correlate strongly with the ratio of external and internal branch lengths. On the other hand, no correlation with distance between sequences in the rRNA trees was observed, thus suggesting that most cases of these alternate topologies represent false-positives due to reconstruction artifacts rather than the accumulation of HGT events with time.

The authors point out that low frequencies of HGT do contradict with the genome content trees (Snel, Bork, and Huynen 1999). They provide explanation for this contradiction.

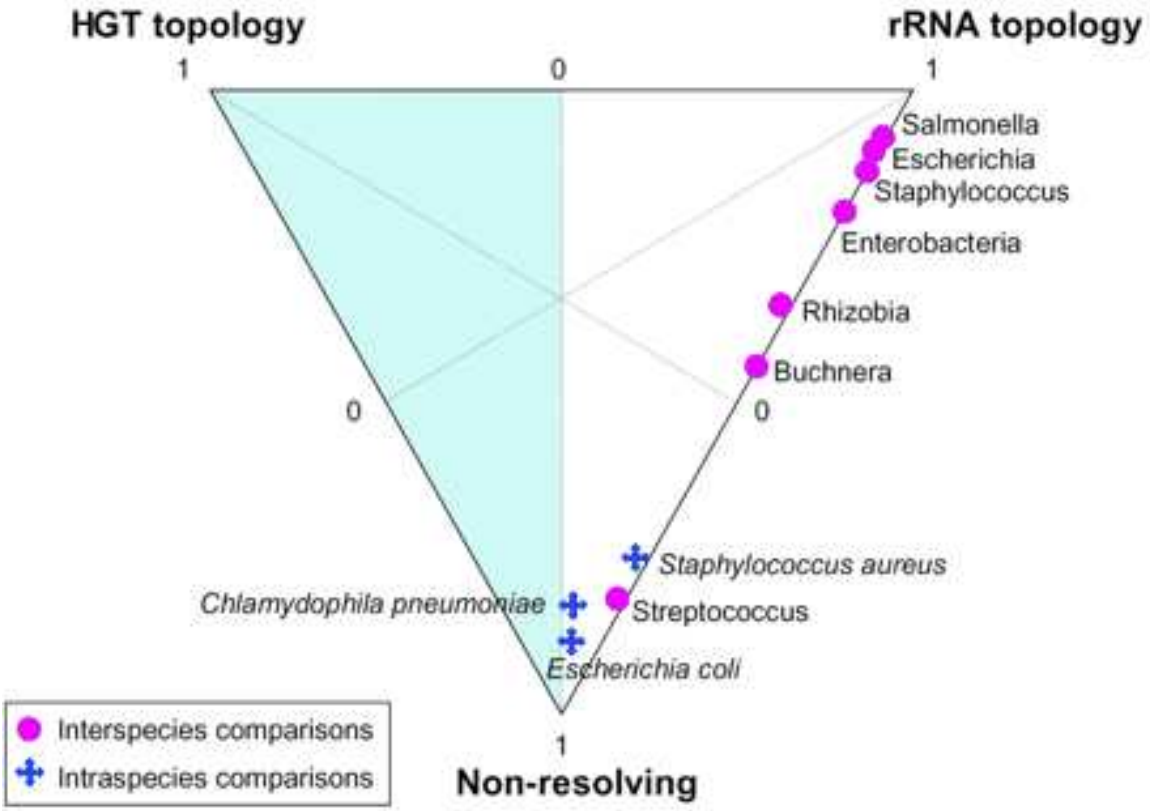


Figure 1.9. Relative frequencies of the three categories of alignments, i.e., those supporting the reference phylogeny (SSU rRNA), those supporting an alternate phylogeny (HGT), and those with no statistical support for any phylogeny. Points represent quartets of genomes for which orthologous genes have been inferred, aligned, and evaluated at the nucleic acid sequences level based on the SH test. The left part of the plot (in blue) represents the area where HGT predominates (from Daubin *et al*).

The genes can be naturally subdivided into two subclasses: those which can be transferred from specie to specie and those genes for which it is possible to find orthologous. In our opinion, this statement is highly questionable, given highly strict criteria for ortholog selection which effectively leaves out many proteins belonging to one or other protein family.

In the recent work Omelchenko *et al.* (Omelchenko *et al.* 2003) studied on more general level putative operons which in part or entirely consist of genes of foreign for a given organism origin.

As the data on operon structure in the majority of genomes are virtually absent Omelchenko *et al.* relied on method of conserved gene pairs (see below) to detect putative operon structures. Phylogenetic trees have been constructed for individual members of operon and compared to the topology of species' tree, as well as to the topologies of phylogenetic trees of other members of putative operon.

The results of this study also do not support SOM: 35 cases of whole operon transfer have been identified in a set of 41 genomes along with 19 cases of mosaic operons, *i.e.* operons containing genes of different phylogenetic origin (which are also more difficult to detect). However, authors note, these results represent conservative low bound estimate for HGT, as very strict cutoffs on ortholog selection and operon detection have been applied.

Regardless of evolutionary premises of operonic organization, it is widely accepted fact that genes in operons are likely to be functionally related. Thus, if operon structure of genome is known, it is relatively straightforward to extract potential functional relations.

The challenge for bioinformatics is that in the majority of sequenced genomes the exact bounds of operons are unknown. This difficulty could have been overcome by exploiting the tendency of genes belonging to the same operon to have smaller intergenic spacers than genes not involved in the same operon. Unfortunately, the situation is complicated by the fact that gene starts often are not accurately predicted by the gene finding algorithms, thus making it difficult to cluster genes into putative operons, judging just by the intergenic distances alone.

Dandekar *et al.* studied coappearance of closely located genes in nine bacterial and archael species (Dandekar *et al.* 1998). Sets of three genomes were selected in such way that fraction of shared orthologs for at least two pairs of genomes in the set was less than 50%. The choice of this threshold has been based on the observation that if percentage of orthologs is less than 50%, the gene order is completely disrupted, meaning that no ancestral colocalization is preserved without evolutionary pressure to maintain such colocalization. In addition, to ensure that conservation of gene order reflects evolutionary constrains rather than horizontal gene transfer events only such genes were considered that were found colocalized in all three genomes.

In each of these three genomes sets authors found approximately 100 genes which were conserved as colinear gene strings. They show that for most of them physical interaction is evident or can be confidently predicted. Most of these proteins participate in core cell activities and represent conserved mechanisms - translation (ribosomal proteins), DNA/RNA-related enzymes and some metabolic enzymes (Fig. 1.10).

It was also observed, that even if in most cases the exact gene order is not conserved, in some cases, operons as unodered sets of genes can be preserved even if extensively rearranged inside. The Trp operon illustrates this point (Fig 1.11). The only aspect of the Trp operon that is preserved in all organisms is the gene pair trpB-trpA, which encodes two subunits of tryptophan synthase.

Authors further speculate that the fact of conservation of the exact gene order in most conserved gene they found suggests the interaction directly after translation or cotranslational folding, which has been experimentally shown for some systems (Netzer and Hartl 1997; Thanaraj and Argos 1996).

Further authors made another important observation: the proteins which are involved in

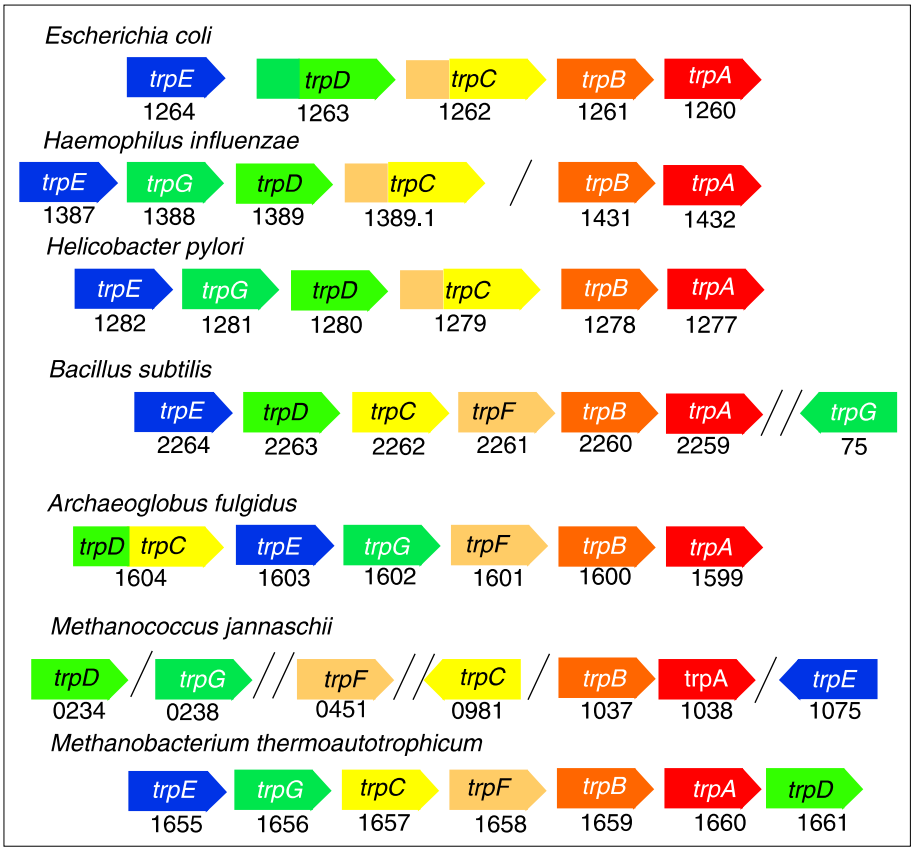


Figure 1.11. Structure of the tryptophan operon in different organisms. Arrows indicate the direction of transcription. Black lines indicate disruption of the operon by intervening genome sequences; double lines indicate a separation of more than 50 genes. The proteins encoded by the genes shown follow: trpA, tryptophan synthase chain; trpB, tryptophan synthase chain; trpC, indol-3-glycerol phosphate synthetase; trpD, anthranilate phosphoribosyltransferase; trpE, anthranilate synthase component I; trpF, anthranilate phosphoribosyl-isomerase; trpG, anthranilate synthase component II. Gene numbers are indicated and are consecutive along the genome. In the proteobacteria, the trpC and trpF genes are fused. The trpG and trpD genes in Escherichia coli, and the trpC and trpD genes in Archaeoglobus fulgidus, are also fused. The only feature of the Trp operon that is conserved across all seven genomes is the trpA–trpB gene pair (from Dandekar *et al.*).

conserved gene pairs are generally more conserved too. For instance, the average degree of sequence identity shared by orthologs that exist as conserved gene pair in *E. coli* and *H. pylori* is 46%; while the equivalent figure for orthologs that do not contribute to conserved gene pairs is 38%. This fact prompted authors to suggest the presence of co-adaption between genes in conserved gene pair. Additionaly, genes which products interact physically should also exhibit a lower rate of evolution. We would like to note though, that significant fraction of the genes found (1.10) are genes coding for ribosomal proteins which belong to the most conserved group of proteins. Thus this result might represent the fact that gene order of ribosomal genes is exceptionally conserved and highest degree of coadaption between parts of ribosome.

Box 2. Confirmed, predicted and putative interactions involving proteins encoded by conserved gene pairs/clusters

Proteobacteria (94 proteins)

(1) Experimentally confirmed (74 proteins).

Ribosomal proteins[17]: Rps9 and Rpl13; initiation factor 3 (IF3), Rpl35 and Rpl20; Rpl21 and Rpl27; elongation factor G (EF-G), Rps7, and Rps12; Rpl7/12 and Rpl10; Rpl1 and Rpl11 are encoded by genes in a large cluster of ribosomal protein genes that includes the gene encoding SecY (L36).

ATP synthase[35]: AtpC, AtpD, AtpG, AtpA and AtpH.

Transporters: ABC transporter subunits[36]; dppB, dppC and dppD dipeptide transporter subunits.

Enzyme pairs/subunits: GroEL and GroES; FrdB and FrdA; NifS and NifU; biotin carboxylase and biotin carboxyl carrier protein; PheT and PheS; ModA and ModB; MraY and MurD; HslV and HslU; ThiD and ThiM; TrpA and TrpB; RpoB and RpoB′; TrpD and TrpE; MreC and MreB.

Regulation: FtsA and FtsZ. The exact ratio is important for division. FtsA acts as a link to the FtsZ ring[37].

(2) Predicted on the basis of experimental data or biological context (18 proteins).

A complex, involving rpl19, RNA methyltransferase and the 21k protein, that participates in ribosome maturation. The 21k protein is in fact a maturase and associates with ribosomal protein[38].

Clp protease (ClpAP) shares structural homology with the proteasome[39], and trigger factor is a prolyl isomerase that could be involved in protein degradation. The existence of the genes encoding these proteins as a conserved pair suggests that trigger factor interacts with ClpAP protease to eliminate misfolded proteins.

A membrane complex formed by glycosylating acyltransferase (LpxA), an acyl carrier protein and three protein-export membrane proteins is functionally plausible and suggested by a conserved gene cluster.

Other examples: CDP ribitol pyrophosphorylase and surface exclusion protein (see text); serine deaminase (SdaA) and the serine transporter (SdaC)[40]; YxjD, YxjE and a short-fatty-acid-chain transmembrane intake protein; the TolB membrane transporter and a peptidoglycan protein in the outer cell wall.

(3) Putative (2 proteins).

NusB and RibE (NusB might in fact facilitate translation of the highly structured *ribE* mRNA).

Gram-positive bacteria (109 proteins)

(1) Experimentally confirmed (83 proteins).

Ribosomal proteins[17]: L11 and L1; S12, S7 and EF-G; a large cluster contains genes that encode SecY and RNA-polymerase-α subunits; L35 and L20; L10 and L7/11; Rps9 and L13; L19 and tRNA methyltransferase; EF-Ts, *mukB* suppressor and ribosome-releasing factor.

ATP synthase: AtpC, AtpD, AtpG, AtpA, AtpH, AtpF, AtpE and AtpB.

Transporters: ATP transporter subunits; fructose permease IIBC component and phosphotransfer protein (in *Bacillus subtilis* the order of transcription is different); oligopeptide permease complex.

Enzyme pairs/subunits: DNA gyrase subunits; RNA polymerase β and β′ subunits; hydroxymethyl-CoA-reductase and pro-lipoprotein diacylglyceryltransferase; thymidilate and folate reductase; pyruvate dehydrogenase; phosphoglycerate kinase and glycerinaldehyde-3-phosphate dehydrogenase; phosphoglycerate mutase and triosephosphate isomerase; pyruvate dehydrogenase; nitrogen-fixation enzymes; DNA helicase; Glu-tRNA amidotransferase (three subunits; the smallest was overlooked in genome sequencing but is also conserved in the cluster); GroEL homologs and GroES homologs.

Regulation: cell-division proteins (two different pairs).

(2) Predicted on the basis of experimental data or biological context (22 proteins).

Phe-tRNA synthetase might interact with IF3. Ribosome interactions similar to those involving Met-tRNA and the ribosome have been measured in initiation complexes[41].

6-Phosphofructokinase and pyruvate kinase. The product of the first enzyme activates the second. Physical coupling would therefore be advantageous.

Heat-shock-stress-response protein BS0069 and the salvage-pathway enzyme hpg transferase might be coupled.

Figure 1.10. From Dandekar *et al, 1998.*

More general method for extraction of conserved gene pairs has been recently proposed

(Overbeek, Fonstein *et al.* 1998; Overbeek, Fonstein *et al.* 1999). It too exploits the observation that some genes, whichever mechanism is responsible for that, often occur in close neighborhood (presumably in the same operon) even in phylogenetically distant species.

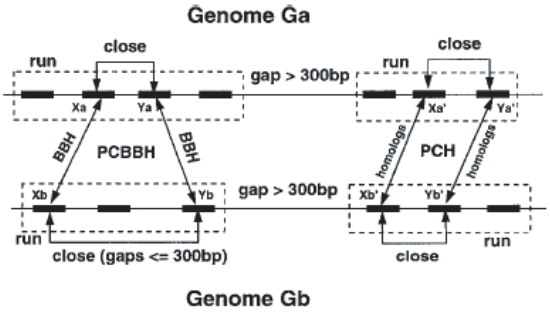


Figure 1.12. Illustration of the definitions of PCBBHs and ‘‘pairs of close homologs’’ (PCHs) (from *Overbeek et al.*).

Let us start with definitions. A "run" in terminology of Overbeek is a set of adjacent genes occurring on a prokaryotic chromosome with maximum size of intergenic spacer smaller than certain threshold D , e.g. 300 b.p. (see Figure 1.12). A pair of two genes X_a and X_b from two genomes G_a and G_b is called bidirectional best hit (BBH) or best-to-best hit if there is recognizable sequence similarity between them lower then certain threshold value P and there is no gene Z_b in genome G_b that is more similar than X_b is to X_a, and there is no gene Z_a in G_a that is more similar than X_a is to X_b. Any pair of genes genes in the run is a "close" pair. A pair of genes from genome G_a (X_a, Y_a) and (X_b, Y_b) from genome G_b form a pair of close bidirectional best hits (PCBBH) if X_a is close to Y_a, X_b close to Y_b, (X_a, X_b) is BBH and (Y_a, Y_b) is BBH (Fig. 1.13). When the bidirectionality of the hit is not required such pair of hits is called simply "pair of close hits" (PCH). Later in the text we will use more general definition "colinear gene pair" or "conserved gene pair" for close pair of homologous genes (supposing we know how to find homologs).

As it was demonstrated in (Overbeek, Fonstein *et al.* 1999): 1) a numerous PCBBHs exist (~60000 in 31 genomes) 2) genes constituting PCBBHs are more likely to be functionally related than can be estimated by random 3) The number of PCBBH is nearly order of magnitude higher than can be found in genomes produced by random shuffling of genes’ locations 4) Almost all PCBBHs are located on the same strand 5) It is possible to reconstruct to a certain extent some core (conserved) metabolic networks 6) PCBBHs are also commonly present in *Archae*.

This approach has become part of WIT[6] system developed at Argonne National Laboratories by Overbeek *et al* (Overbeek, Larsen *et al.* 2000).

[5] Available on-line at http://wit.mcs.anl.gov/WIT2/.

This work has been extended to a more general approach in STRING[7] (Snel, Lehmann *et al.* 2000). Instead of considering only pairs, colinear strings are sought for. Moreover, iterative extraction of such strings is implemented. That is, for gene X found in string S, all colinear strings S_i X belongs to are found, then for gene Y neighboring X in the string S_k, all strings S_j containing Y are found; and so on. This iterative search can not be continued *ad infinitum* though, because generally it does not converge and only after few iterations results in combinatorial explosion.

In more recent work (Mering1 *et al.* 2003) STRING has been extended to use gene fusion and phylogenetic profiles along with conserved colinear strings to predict functional relations. The scoring system has been devised which would allow for combined score of functional relatedness of two genes based on all three approaches.

In related work Wolf *et al.* (Wolf *et al.* 2001) introduced entirely different method of extracting colinear gene strings. The method is based on idea of genome order alignment, similar to the sequence alignment but instead considers genes as basic informational characters (by using COGs). The method has been shown to detected conserved operons such as ribosomal operons. Although the method is interesting and applicable, it has one serious drawback - it's sequence-like alignment would totally miss shuffled colinear gene strings, e.g. identical operons in different genomes containing the same genes but in varying order.

Despite of success of colinear gene strings method it does not approach the problem mentioned in introductory section of this work. As it has been shown in (Mushegian and Koonin 1996) and (Itoh *et al.* 1999) operons are generally poorly conserved. Consequently only limited amount of functionally coupled genes can be detected by using colinear strings approach. Only in the case when certain constraints on genes' colocalization are present such colinearity would be observed in sufficient amount of species. Furthermore, as it has been mentioned earlier in this chapter, even the proteins involved in the fusion events, which by itself implies tight physical interaction, are often located in distant regions on chromosome. What if there are no constraints on genes colocalizaton? Is it still possible to exploit non-randomness of gene order in prokaryotes?

We would like to highlight again two simple facts we know about operons: i) genes in operons are likely to be functionally related and ii) operons are not conserved, *i.e.* the gene content of operons always varies from specie to specie. Two works has been published recently aiming at this contradiction: (Lathe III *et al.* 2000) and work of our group (Kolesov *et al.* 2001) which will be discussed in details in the following chapter.

To explain aforementioned difficulty Lathe III *et al* put forward the concept of *uber-operon*. Uber-operon is set of functionally related genes encompassing several operons. In the course of evolution genes within this set are randomly sampled, constituting to the random subsets of uber-operon - the operons. In this scenario operons would appear different in each genome, although the genes composing them would still be functionally related to each other.

[6] available on-line at http://bork.heidelberg.de/strings

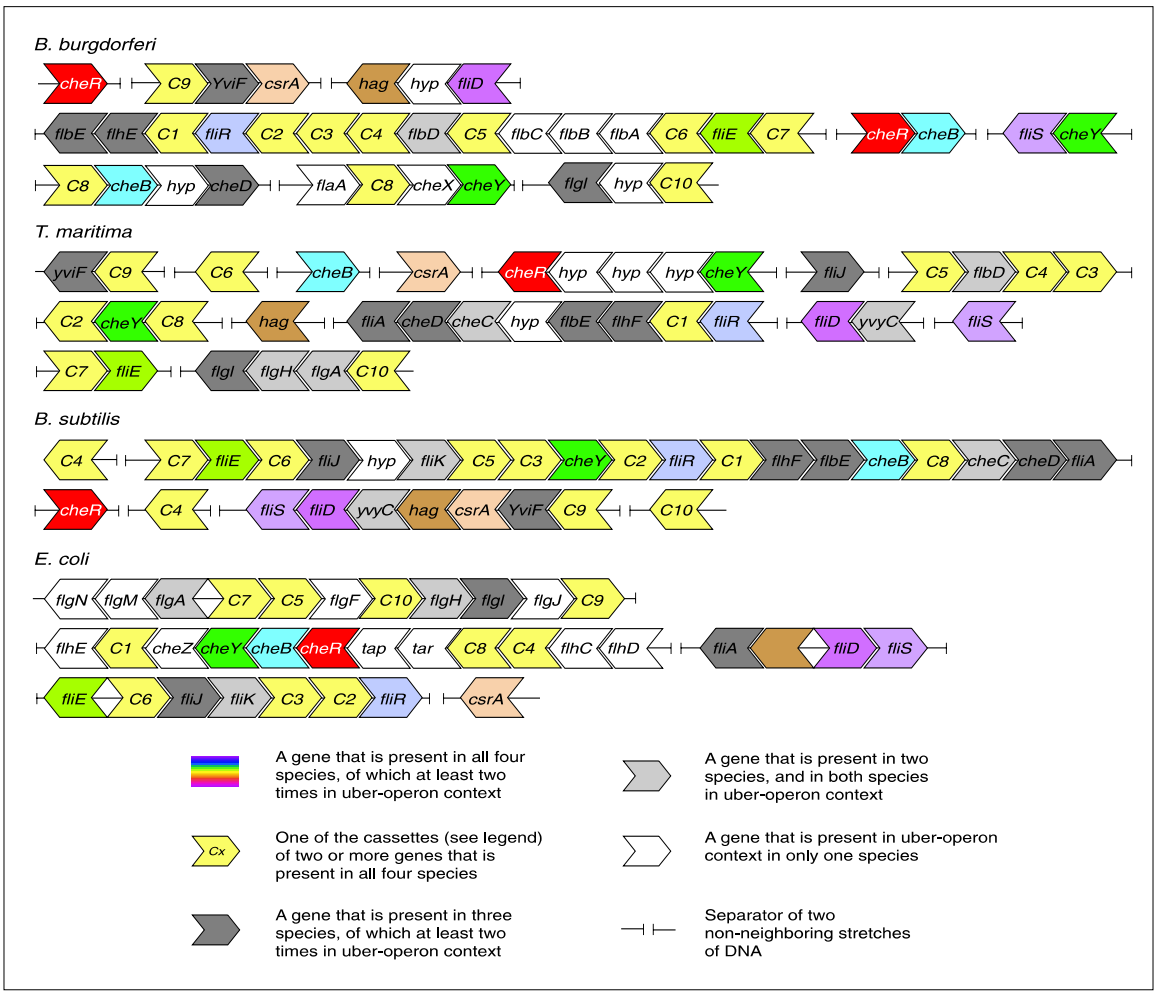


Figure 1.13. The genomic organization for four species of (i) the genes from the flagellar uber-operon (colored), (ii) genes that occur in the context of the uber-operon but not in all species (gray) and (iii) genes that happen to occur once in the context of the uber-operon (white). The white genes are mostly absent from complete genomes of the other species. The composition of the cassettes is (transcription direction from left to right): C1, (flhB, flhA); C2, (fliO, fliP, fliQ); C3, (fliL, fliM, fliN); C4, (motA, motB); C5, (flgE, flgD); C6, (fliF, fliG, fliH, fliI); C7, (flgB, flgC); C8, (cheW, cheA); C9, (flgK, flgL); C10, (flhO, flgG). The white genes labeled 'hyp' are hypothetical (genes with unknown function) and are not homologous to each other.

Indeed, few uber-operons have been found by manually iterating through orthologs of relevant genes and their neighbors. Example of flagellar-related genes uber-operon is shown on Figure 1.13.

Although the model of uber-operons is in the good agreement with the observed gene shuffling in operons, there are several difficulties which hinder automatic extraction of uber-operons: i) not necessarily all genes in operon are functionally related - although the genes in operon are likely to be functionally related some of the neighbors can be

completely random; ii) the exact bounds of operons are often unknown[8] , leading to uncertainty in deriving 'true' operon neighbors. Both of these problems lead to a highly undesirable effect: even one 'wrong', i.e. functionally unrelated to its neighbors gene can cause our resulting uber-operon be badly polluted with false positives if simple algorithm of iterative extraction of neighbors is being used. Hence, another, more sophisticated, aware of these difficulties approach is required.

[7] At the time of writing the only moderately complete databases of experimentally confirmed operons are available for *E. coli* and *Bacillus subtilis*.

Chapter 2

SNAPping up functionally related genes in prokaryotic genomes

In this chapter we will discuss our own approach to finding functionally related genes using properties of gene order in prokaryotes without relying in any way on it's conservation.

2.5. Main ideas and definitions

Genes fulfilling the same function in different organisms, or similar but distinct functions in the same organism are expected to possess a certain degree of sequence similarity due to the evolutionary conservation of their primary structure. By contrast, functionally related genes are essentially different genes that are involved, for example, in the same metabolic or signaling pathway. Such genes are normally not similar; hence, their relatedness is not detectable by sequence comparison. Instead, functionally related genes often form clusters on the chromosome (Lawrence and Roth 1996); their relatedness may be manifested by spatial proximity rather than structural resemblance. Throughout this text, we will use the terms S-relationship, N-relationship, and SN-relationship to describe the cases where genes are related by similarity, neighborhood, or a mixture thereof, respectively.

In this work we attempt to exploit the observation that neighboring genes on bacterial chromosomes tend to be functionally related, even if there is no evidence that their positional preference with respect to each other is conserved across many different genomes. Potentially, any random pair of adjacent genes could be functionally coupled. It is evident, of course, that many hundreds and even thousands of genes encoded in complete bacterial genomes fall into hundreds of different functional categories, making the joint occurrence of two functionally related genes a rather unlikely event (Huynen *et al.* 2000). We need to be able to distinguish random pairs of physically proximate genes from meaningful ones, without relying, in general, on the conservation of such pairs across multiple genomes.

Before we provide a formal description of our algorithm, we start with a simple illustration. Let us first consider a group of five genes involved in a certain biochemical process, and compare this group as a whole with functionally related groups in other genomes. In the case of a perfectly conserved gene cluster, we will observe a string of genes α_A,β_A,γ_A,ω_A,ε_A in the genome A, α_B,β_B,γ_B,ω_B,ε_B in the genome B, α_C,β_C,γ_C,ω_C,ε_C in the genome C, and so on, such that the genes from different genomes denoted with the same Greek letter are S-related, and the genes from the same genome are N-related. In a more complex, and more realistic case, many of the inter-genome S-relationships may not be preserved due to physiological differences between the species involved, or simply

because the similarity is not detectable with current sequence comparison tools. Likewise, and even more probably, the N-relationships within each genome may be disrupted as a result of gene shuffling in the course of evolution. Therefore, the association between the different instances of this particular gene cluster in different genomes will be expressed as an irregular mixture of S- and N-relationships.

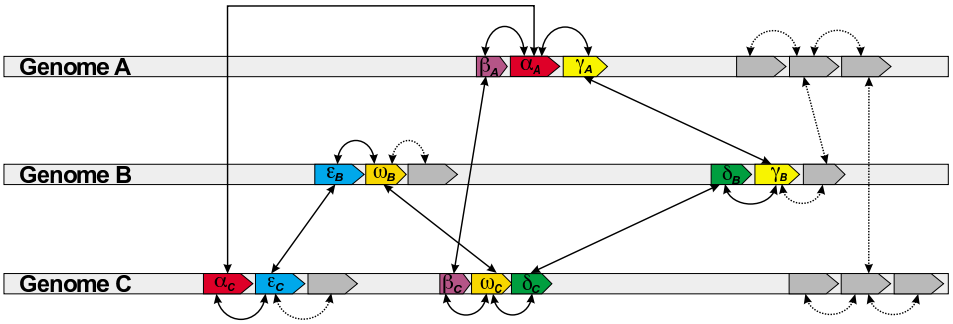


Figure 2.1. Finding genes functionally coupled with the gene α residing in the genome A. Colored arrows represent individual genes and their direction. Straight black arrows represent S-relationships between orthologs in different genomes while round black arrows represent N-relationships between genes in the same genome. Only one neighbor of every gene in each direction is considered. The analysis starts with finding neigbours of gene α genes β_A and γ_A in the genome A. Then their orthologs on other genomes are identified, and so on. As a result, a chain of alternating similarity- and neighborhood relationships, called SN-graph, is constructed. In this example, the SN-graph has a closed path $\alpha_A, \gamma_A, \gamma_B, \delta_B, \delta_C, \omega_C, \omega_B, \varepsilon_B, \varepsilon_C, \alpha_C, \alpha_A$ or SN-cycle, indicating that at least some part of the constituent genes may be functionally related. Solid black arrows correspond to the closed path while the rest of the SN-graph is shown in dotted arrows. Genes not participating in the closed path are shown in grey.

Let us consider a hypothetical example depicted in Figure 2.1 and focus on the chain of SN-relationships originating from gene α in genome A. This gene is N-related to the genes β_A and γ_A. Gene γ_A is S-related to γ_B, the latter is N-related to δ_B, and so on. The complete system of such SN-relationships, subject to certain limitations described below, forms an *SN-graph*. SN-paths on the graph are made up of alternating S- and N-relationships. The former are derived using selective sequence comparison tools, such as BLAST, (Altschul *et al.* 1997) and are thus extremely significant. By contrast, the latter are overwhelmingly random. For this reason, the majority of the SN-paths has no diagnostic value. However, intermixed with a large number of "false positives" among N-relationships, i.e. pairs of totally unrelated genes, are a number of N-related genes that are actually functionally coupled. We put forward a hypothesis that such meaningful N-relationships are likely to occur in closed SN-paths, which we will call SN-cycles. In Figure 2.1, the longest SN-cycle is represented by the path $\alpha_A, \gamma_A, \gamma_B, \delta_B, \delta_C, \omega_C, \omega_B, \varepsilon_B, \varepsilon_C, \alpha_C, \alpha_A$. The primary intuition here is that the N-relationships resulting from nonrandom associations between genes will have a statistical tendency to throw a bridge between pairs of S-related proteins, and ultimately help join proteins that belong to the same metabolic

pathway, resulting in a closed path on the graph. Our principal approach in this work is to exploit simultaneously the two possible types of relatedness between genes - S- and N-relationships - in order to establish functional links undetectable by either type of relationship alone.

2.6. Description of the algorithm

We consider N bacterial genomes $G_i (i = 1, N)$, each containing M^i genes g_k^i $(k = 1, M^i)$, where k is the sequential number of the gene on the chromosome. Two genes g_k^i and g_{k+1}^i from the same genome i are N-related if they fulfill the following conditions: i) both genes g_k^i and g_{k+1}^i have the same orientation (i.e., are situated on the same strand; as demonstrated by (Overbeek, Fonstein *et al.* 1999), co-occurrence of functionally coupled genes on opposite strands is a very rare event), and ii) the distance between the stop codon of 5' gene (g_k^i if on direct strand, g_{k+1}^i if on reverse) and the start codon of 3' (g_{k+1}^i on direct strand) is smaller then a certain threshold value d (typically 500 base pairs.). We take into account spatial association between genes that are at most c genes away from each other. Therefore, a genome i can be represented as an unordered set of up to $M^i - 2c$ gene words,$W_q^i, (q = c+1, M^i - c)$, each word being an ordered list of up to $2c + 1$ genes:

$W_{c+1}^i = (g_1^i, \ldots, g_{2c+1}^i), W_{c+2}^i = (g_2^i, \ldots, g_{2c+2}^i), W_{c+3}^i = (g_3^i, \ldots, g_{2c+3}^i)$, etc.

In other words, each gene word W_q^i contains the gene g_q^i, its c neighbors on the left, and its c neighbors on the right. A genome will contain exactly $M^i - 2c$ gene words only if all genes are on the same strand and are separated by no more than d bases. Since this is never the case, the actual number of gene words in a genome will be smaller. For the same reason many of the gene words will contain less than $2c+1$ genes. The minimal number of genes in a gene word is 2 since otherwise no N-relationship in the word can exist. Throughout his work we used $c = 2$ (unless otherwise stated) in order to make our tests computationally feasible.

An all-against-all comparison of the genes , $(i = 1, N, k = 1, M^i)$ is conducted using the PSI-BLAST algorithm (Altschul *et al.* 1997). An S-relationship between two genes g_k^i and , residing on the genomes G_i and G_j, respectively, exists if the BLAST E-value $E(g_k^i, g_l^j) < e$ and the coverage of the BLAST alignment, defined as the fraction of amino acids of the shorter compared protein covered by the alignment, $C(g_k^i, g_l^j) > a$ where e and a are parameters of the analysis. As an additional restriction, we may require the BLAST match to be reciprocal, such that $E(g_k^i, g_l^j) < e$, $E(g_l^j, g_k^i) < e$ and there is no $x = 1, M^i, x \neq k$ and $y = 1, M^j, y \neq l$ such that $E(g_x^i, g_l^j) < E(g_k^i, g_l^j)$ and $E(g_y^j, g_k^i) < E(g_l^j, g_k^i)$ The matrix of all-against-all BLAST matches is made symmetrical by selecting for each pair of proteins the best E-value and the best value of coverage C, such that

$E'(g_k^i, g_l^j) = E'(g_l^j, g_k^i) = min(E(g_k^i, g_l^j), E(g_l^j, g_k^i))$ and
$C'(g_k^i, g_l^j) = C'(g_l^j, g_k^i) = max(C(g_k^i, g_l^j), C(g_l^j, g_k^i))$

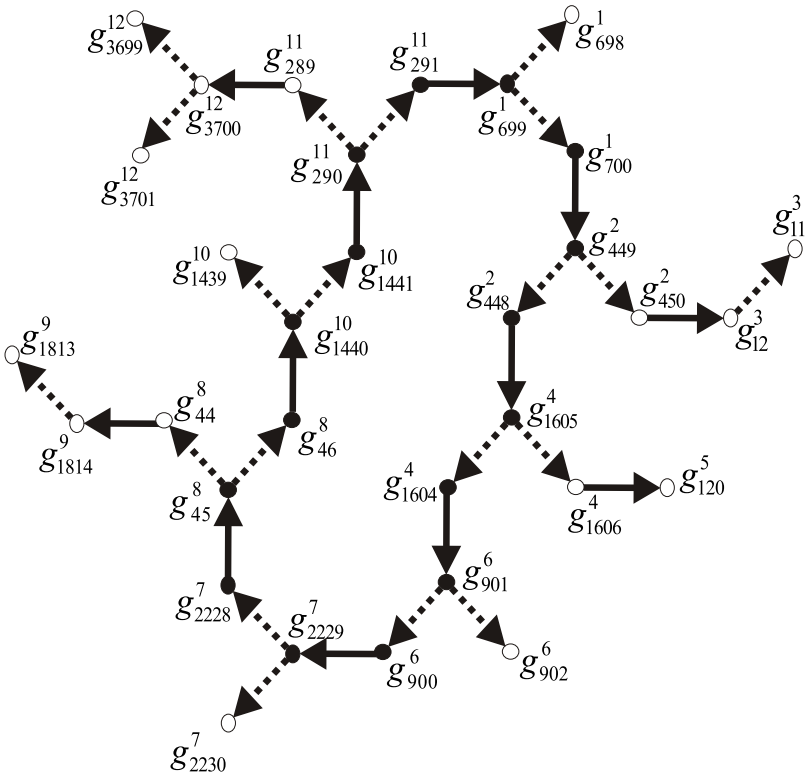


Figure 2.2. A hypothetical chain of SN-relationships. A part of a hypothetical SN-graph involving an SN-cycle. Genes participating and not participating in the SN-cycle are shown as filled and open circles, respectively, and are denoted as where the superscript stands for the genome number and the subscript for the sequential gene number on the chromosome. Solid and dashed arrows depict similarity and neighborhood relationships, respectively. The number of gene neighbors considered on each side $c = 1$.

We can now represent the chain of SN-relationships originating from an arbitrary gene as an SN-graph involving S- and N-relationships in an alternating fashion, starting either with an S-relationship or an N-relationship in which is involved. An example of such a graph is shown in Figure 2.2. It is easy to see that the SN-graph joins gene words that have at least one pair of S-related genes.

In our implementation, an SN-graph is traversed using the depth-first algorithm and all closed SN-paths, or SN-cycles, are identified. In Figure 2.2, an SN-cycle involves 16 genes shown as filled circles, corresponding to the 8 related gene words. A special case of an SN-cycle is constituted by colinear gene clusters in which the order of genes is partially or fully conserved across several genomes. Such SN-cycles involve words with more than one pair of S-related genes (Figure 2.3).

With an increasing number of genomes the number of nodes in the SN-graph grows very quickly so that finding all paths becomes computationally prohibitive. To demonstrate the feasibility of our approach, without losing the generality, we set an upper limit on the path length at a certain value, typically 14 nodes.

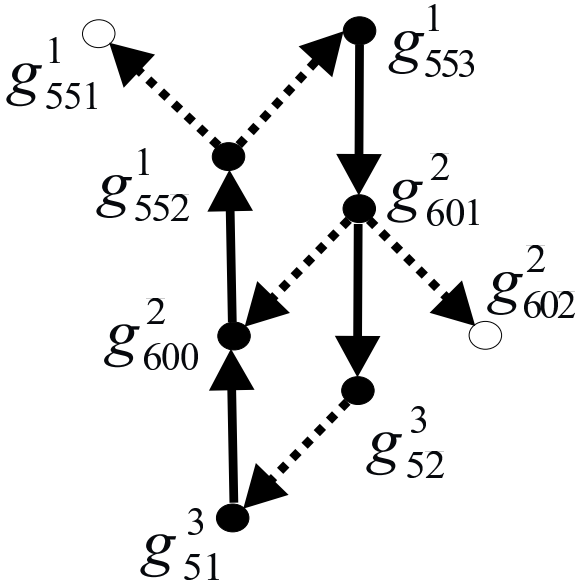


Figure 2.3. A hypothetical SN-graph which involves a conserved pair of genes in three genomes: genes 552 and 553 in genome 1, genes 600 and 601 in genome 2, and genes 51 and 52 in genome 3. In this case, the SN cycle is equivalent to a colinear gene cluster of the type described by Overbeek *et al.* (1999). Notation as in Figure 2.2.

2.7. Measuring the performance of the method

All genes belonging to an SN-cycle are regarded as functionally coupled. In order to test the validity of this assertion, we need to measure the performance of the algorithm on a large number of documented cases of "true" functional relatedness. Two different approaches for defining the standard of truth for our calculations have been explored.

2.7.1. Analysis of reference metabolic pathways

The entire KEGG/PATHWAY database[9] (Kanehisa and Goto 2000) was downloaded from ftp://kegg.genome.ad.jp. The database was processed with a sophisticated *Perl* script to extract the pathway graph in a form suitable for subsequent computer analysis. Information about links between biological objects cannot be easily gleaned from the KEGG image files representing the pathways. We obtained this information indirectly by comparing the list of all biochemical reactions present in the database with another list which specifies both the EC number of a given enzyme and the compounds it interacts with. Since the names of the compounds in the first and the second list are often inconsistent, we used a sub-string comparison technique to establish correspondence between them. Further, unspecific widely applicable metabolites, such as H_2O, alcohol, CO_2, etc. were not considered.

The pathway graph is constituted by vertices and edges corresponding to enzymes and

[8] available on-line at http://www.genome.ad.jp/kegg/kegg2.html

substrates, respectively. Given a set of enzymes represented by their EC numbers $E = (E_1, E_2, \ldots, E_n)$, where n is the number of enzymes in the set, our goal is to find a measure, $0 \le K_p \le 1$, to describe their "concentration" on the pathway graph. We call this measure "pathway coefficient". The ideal case of $K_p = 1$ corresponds to an SN-cycle joining enzymes that form a compact pathway sub-graph such that i) no other nodes except for $(E_1, E_2, \ldots, E_n)$ exist, and ii) for any nodes E_i and E_j there exists a path connecting them. The worst case $K_p = 0$ describes an SN-cycle that joins totally unrelated enzymes, i.e. there is no path on the pathway graph connecting any pair of the enzymes found.

The metabolic distance D_{ij} between two enzymes E_i and E_j on the pathway graph is defined as the minimal number of reaction stages (edges) connecting these enzymes (vertices). Given a set of enzymes, we used the following approach to determine the value of the pathway coefficient K_p. Single linkage clustering was applied to the metabolic distance matrix D_{ij}, $i = 1, n$, $j = 1, n$ in order to find the largest cluster of vertices $C \in E$ subject to the constraint that $D_{ij} < D_t$, where D_t is the threshold metabolic distance. The pathway coefficient can then be computed as:

$$K_p = \lambda_p \frac{m}{n}$$

where m is the number of elements in C, and λ_p is a normalization coefficient defined as:

$$\lambda_p = \frac{m}{\sum_{j=1}^{m} q_j}$$

where q_j denotes the number of times the EC number corresponding to the jth element of C occurred in the entire pathway graph.

2.7.2. Utilization of functional categories

The degree of functional coupling between the genes involved in SN-cycles was also examined in reference to the MIPS functional role catalogue [10] developed for the yeast genome (Mewes *et al.* 1997). The catalogue has a hierarchical structure. Each of the 15 main classes (e.g. metabolism, energy etc.) contains three to four subclasses, with the total number of functional categories exceeding 200. Correspondingly, the numeric designator of a functional class can include up to four numbers. For example, the yeast gene product YGL237c is attributed to the functional category 04.05.01.04, where the numbers, from left to right, mean transcription, mRNA transcription, mRNA synthesis, and transcriptional control. Nearly 4000 yeast genes could be ascribed to at least one functional category based on careful manual analysis of extrinsic evidence (similarity to known proteins, presence of indicative sequence patterns) as well as experimental data from the literature. In this work, the MIPS classification was used for automatic assignment of functional categories to gene products from completely sequenced genomes based on significant homology to one or many functionally characterized yeast genes. The functional category coefficient for a group of genes with at least one functional category assigned $F = (F_1, F_2, \ldots, F_n)$ was computed as:

[9] Available on-line at http://mips.gsf.de/proj/yeast/catalogues/funcat/index.html

$$K_f = \lambda_f \frac{m}{n}$$

where n is the number of genes in the group, m is the maximal number of times a functional category f occurred in F, and λ_f is a normalization coefficient:

$$\lambda_f = 1 - P(m, f)$$

In the latter equation $P(m, f)$ denotes the binomial probability of the functional category f to occur m times in the group of genes of size n:

$$P(m, f) = \frac{n!}{m!(n-m)!} p_f^n (1 - p_f)^{n-m}$$

where the probability p_f is estimated as the general frequency of occurrence of a functional category f.

2.7.3. Implementation and data sources

The main vehicle for the present study was the PEDANT genome analysis system (Frishman and Mewes 1997; Frishman, Albermann *et al.* 2000). The PEDANT database[11] contains exhaustive functional and structural annotation of all completely sequenced genomes. In particular, gene products are automatically assigned to yeast functional categories (Mewes *et al.* 1997) and enzyme classes (Kanehisa and Goto 2000) based on similarity searches. Out of 35 finished genomic sequences available at the time of writing, we selected 12 genomes from sufficiently distant species, as assessed visually based on a maximum likelihood phylogenetic tree derived from the small-subunit rRNA sequences using the PHYLIP package (Felsenstein 1989). Namely, these genomes are: *Aeropyrum pernix, C. jejuni, C. pneumoniae, E. coli, M. pneumoniae, M. thermoautotrophicum, Mycobacterium tuberculosis, Pyrococcus abyssi, T. acidophilum, T. maritima, Synechocystis sp.* and *T. pallidum*.[12] Throughout this text, gene IDs as available through the PEDANT database are utilized.

A *Perl* program was written to extract gene positional information and various other attributes from the PEDANT MySQL relational tables, build the SN-graphs, detect SN-cycles, and study the features of the genes predicted to be functionally related.[13]

2.8. Results

2.8.1. Formal properties of SN-cycles

We begin with asking two questions: (i) do non-trivial SN-cycles (i.e. those not involving colinear gene clusters) exist; and (ii) if they exist, what is the chance that they occur at random. To answer the first question, it is sufficient to provide an example. Figure 2.4 shows a closed system of SN-relationships involving some of the genes responsible for lysine biosynthesis in the prokaryotes. There are three adjoining SN-cycles originating at

[10] Available on-line at http://pedant.gsf.de

[11] URLS of the respective sequencing centers are available at http://pedant.gsf.de/credits.html

[12] Later the algorithm was re-implemented for performance (see Chapter 3).

the *E. coli* coli gene coding for dihydrodipicolinate reductase. The detailed discussion of this example from the functional point of view will follow later. In order to answer the second question, we have studied the behavior of SN-graphs and their dependence on various analysis parameters using a set of 12 completely sequenced genomes from phylogenetically distant species (see 2.7.3). Figure 2.5 shows the dependence of the number of SN-cycles identified from the number of genomes used in the analysis. The graph makes immediately obvious the value of a large number of sequenced genomes in comparative genomics: there is a boost in the number of SN-cycles found as the number of genomes approaches ten. This is in agreement with the results of Overbeek *et al.*, who noted that in order to detect functional coupling for a given functional subsystem, at least ten genomes are needed.

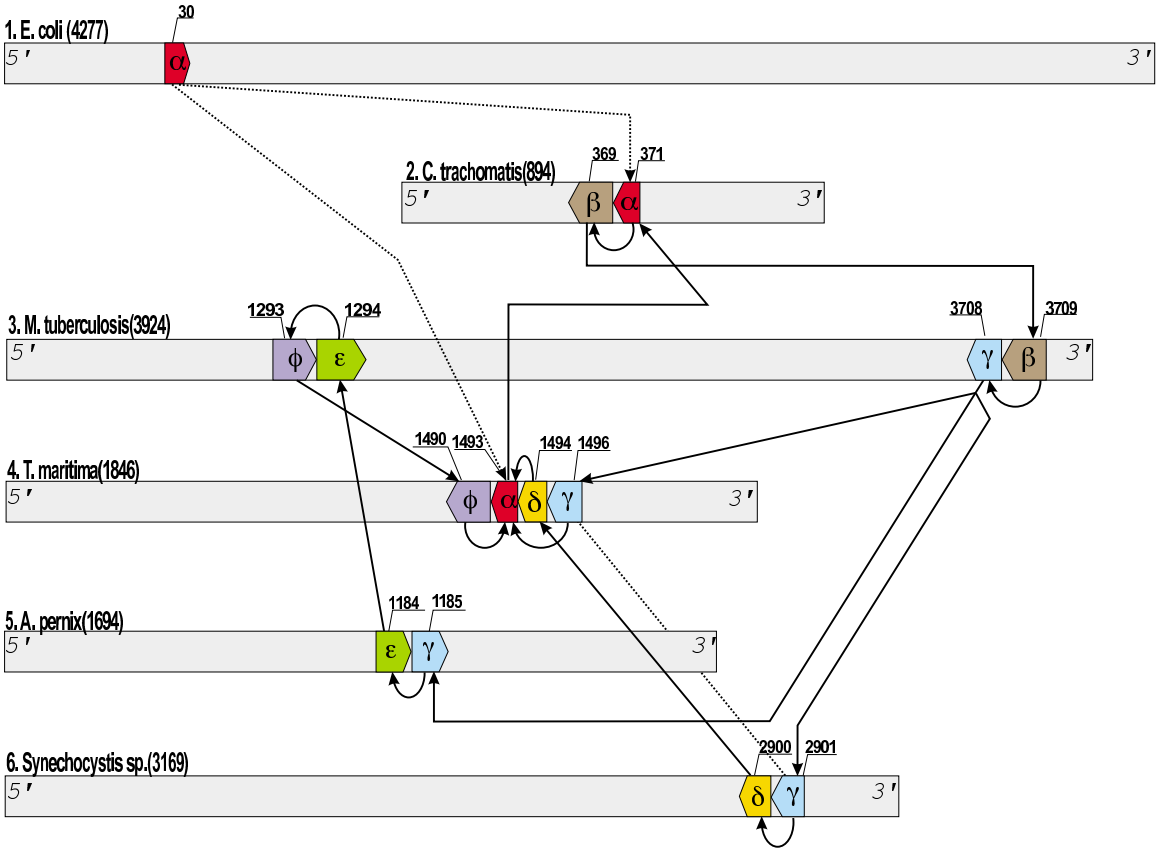
The same experiment was performed with our set of 12 genomes after randomly shuffling the gene order within each genome, which effectively leads to destroying meaningful N-relationships while keeping S-relationships intact. The difference in the occurrence of SN-cycles in real and shuffled genomes quickly grows with the number of genomes and becomes especially pronounced when more than ten genomes are considered. In the complete set of 12 genomes with real gene order, 33,000 SN-cycles were found, as opposed to 3500 SN-cycles in shuffled genomes. It should also be noted that at greater evolutionary distances between species, the share of non-random SN-cycles increases. We thus estimate that with a sufficiently large number of evolutionary distant genomes taken into account, approximately 90% of SN-cycles are non-random. Moreover, as seen in Figure 2.5(b), the increase in the number of SNcycles is almost exclusively caused by long (more than ten nodes) SN-cycles. Due to the virtual disappearance of long SN-cycles after shuffling, we are compelled to conclude that the majority of all such cycles reflect conserved spatial association between genes, although certain parts of these cycles may still be random. As expected, detection of SN-cycles is strongly influenced by the choice of the BLAST alignment parameters (Figure 2.5(c) and (d)); their number grows quickly as the BLAST parameters are changed from very stringent (E-values close to 0, coverage close to 100%) to entirely permissive (any E-value, any coverage). However, even with the most permissive parameters, the number of SN-cycles identified in real, unshuffled genomes is nearly an order of magnitude higher than in the genomes with random gene order.

Since the S-relationships are not influenced by gene order shuffling, the difference observed is solely due to the strong functional coupling of adjacent genes in the former and the virtual disappearance of the N-relationships in the latter.

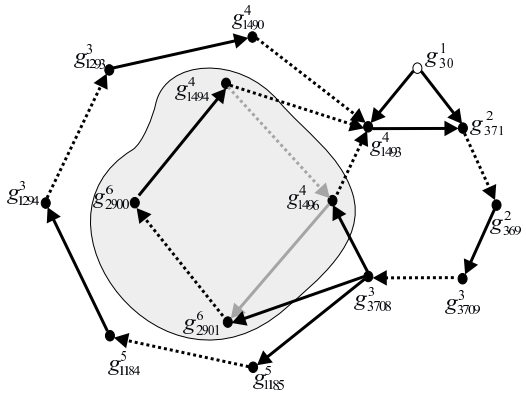
2.8.2. Functional content of SN-cycles

Now that we have formally established the overwhelming non-randomness of long SN-cycles and their frequent occurrence, it is time to examine their functional content. The central issue in accessing the performance of our method is the granularity of the functional assignments. Similarity-free approaches are necessarily less specific than methods based on protein sequence and structure comparison. While the latter are often capable of

(a)


1. E. coli (4277)
30
5′
α
3′
2. C. trachomatis(894)
369
371
5′
β
α
3′
3. M. tuberculosis(3924)
1293
1294
3708
3709
5′
ϕ
ε
γ
β
3′
4. T. maritima(1846)
1490
1493
1494
1496
5′
ϕ
α
δ
γ
3′
5. A. pernix(1694)
1184
1185
5′
ε
γ
3′
6. Synechocystis sp.(3169)
2900
2901
5′
δ
γ
3′

(b)


g^4_{1490}
g^3_{1293}
g^1_{30}
g^4_{1494}
g^2_{371}
g^4_{1493}
g^3_{1294}
g^6_{2900}
g^4_{1496}
g^2_{369}
g^3_{3708}
g^3_{3709}
g^6_{2901}
g^5_{1184}
g^5_{1185}

(c)

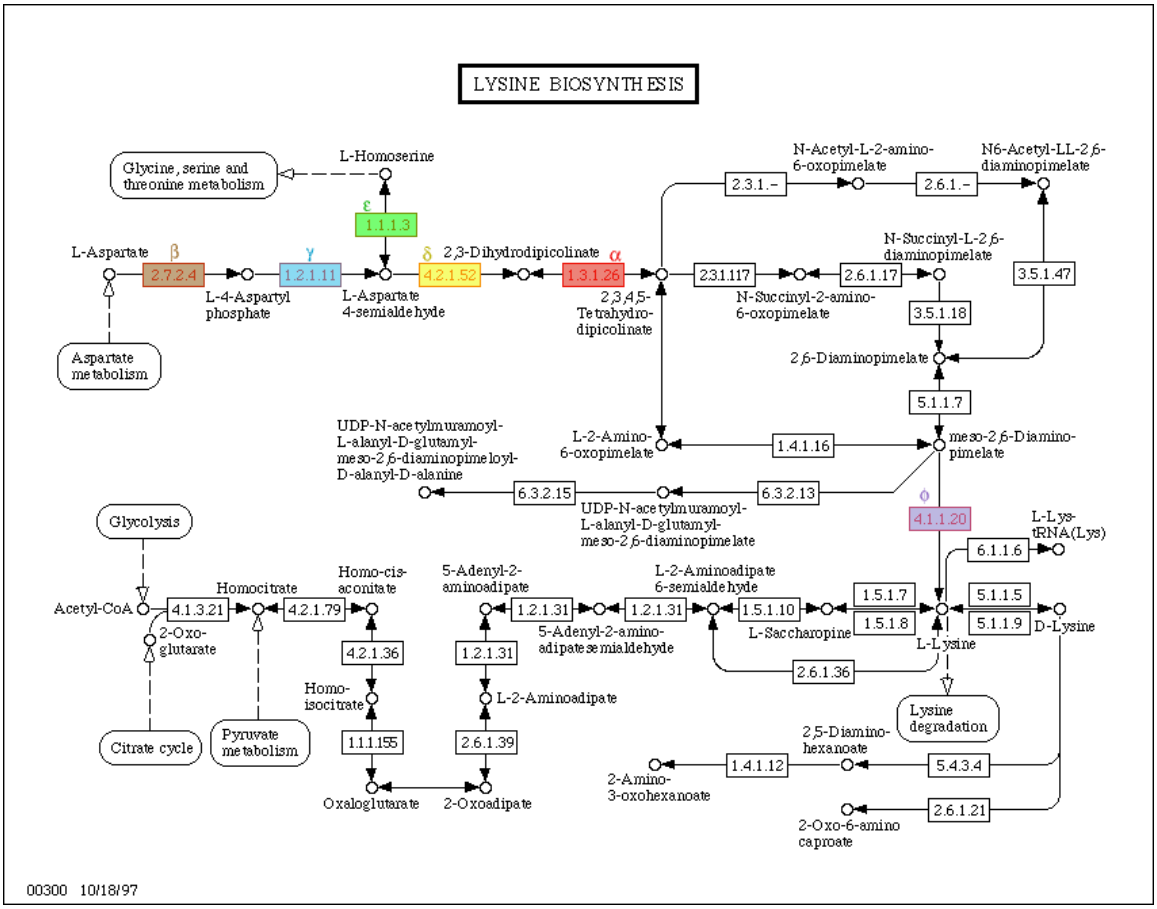


Figure 2.4. SNAP analysis of the *E. coli* gene g1786214 coding for dihydrodipicolinate reductase. A part of the SN-system originating from the *Chlamydia trachomatis* gene gi_3328787 (which is orthologous to the *E. coli* gene) is shown. For illustration purposes, only six prokaryotic genomes are considered, numbered from 1 to 6. (a). A representation of the gene location and their S and N- relationships. The total number of genes in each genome is shown in parentheses. Sequential numbers of genes, counting from the 5' to the 3' end of the genome are indicated. Additionally, each gene is colored and labeled with a Greek letter according to its function: α (red), dihydrodipicolinate reductase (EC 1.3.1.26); β (brown), aspartokinase (EC 2.7.2.4); γ (cyan), aspartate-semialdehyde dehydrogenase (1.2.1.11); δ (yellow), dihydrodipicolinate synthase (EC 4.2.1.52); ε (green), homoserine dehydrogenase (EC 1.1.1.3); ϕ (lilac), diaminopimelate decarboxylase (EC 4.1.1.20). Three adjoining SN-cycles are present: (i) g^2_{371} g^2_{369} g^3_{3709} g^3_{3708} g^4_{1496} g^4_{1493} ;(ii) g^2_{371} g^2_{369} g^3_{3709} g^3_{3708} g^5_{1185} g^5_{1184} g^3_{1294} g^3_{1293} g^4_{1490} g^4_{1493} ; and (iii) g^2_{371} g^2_{369} g^3_{3709} g^3_{3708} g^3_{3708} g^6_{2901} g^6_{2900} g^4_{1494} g^4_{1493}.Incidentally, a simple colinear gene cluster involving the spatially conserved pair of genes b and g in T. maritima and Synechocystis sp. is present; the extra S-relationship between the genes of the type γ is shown as a broken line. (b) An SN-graph corre- sponding to the system shown in (a). The shadowed part of the graph stems from the conserved pair of adjacent genes that have sequential numbers 1494 and 1496 in the genome of T. maritima and number 2900 and 2901 in the genome of Synechosistis sp.

(c) A part of the KEGG metabolic map involving the six genes predicted to be functionally coupled. Enzymes (highlighted in the same colors as used in (a)) encoded by the genes α, β, γ, δ and ε catalyze subsequent reactions in the lysine biosynthesis pathway, while the reaction catalyzed by the enzyme ϕ is separated from the nearest reaction of the first group by two other metabolic steps.

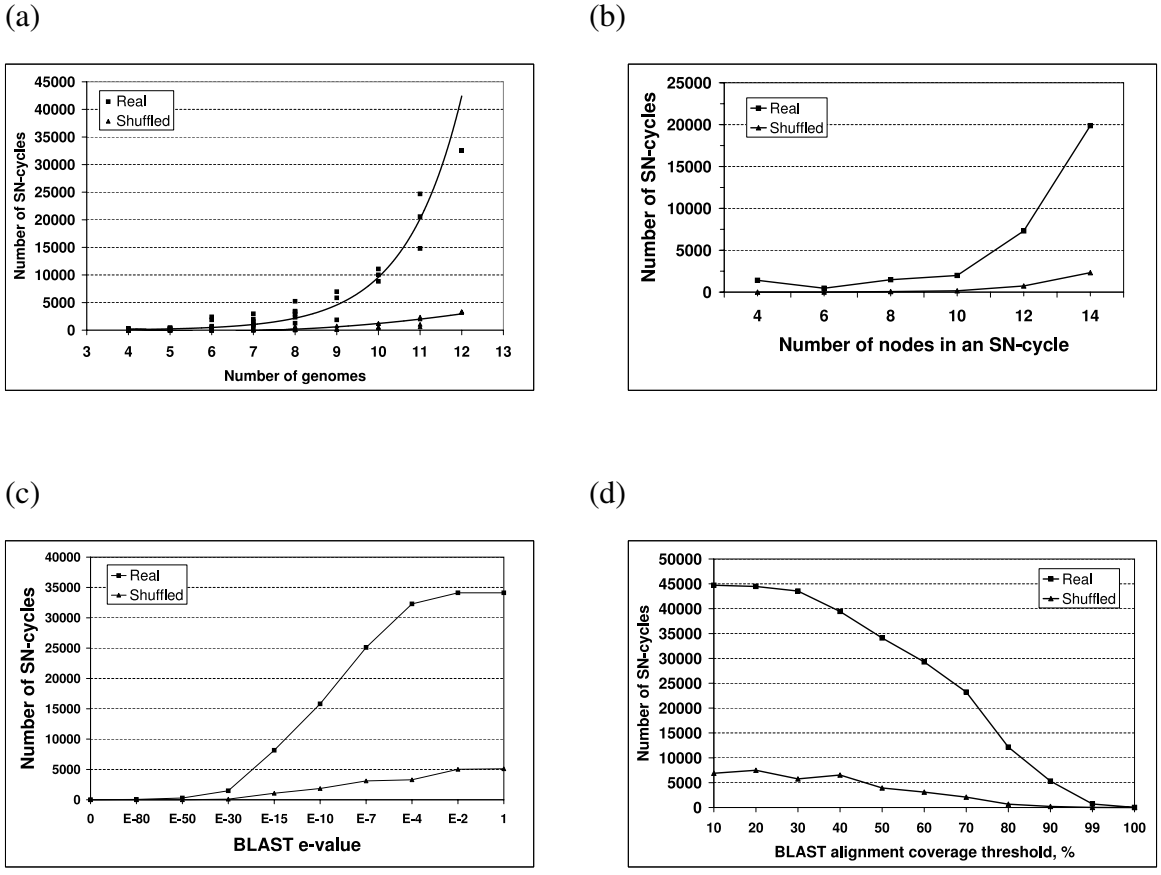


Figure 2.5. Comparison of the global properties of SN-cycles in real (squares) and shuffled (triangles) genomes. Dependence of the number of SN-cycles detected on (a) the number of genomes considered (in order to make computations feasible, only selected data points were computed), (b) cycle length, (c) BLAST cutoff E-value, and (d) BLAST alignment coverage is shown. The default parameters, unless explicitly specifed are: BLAST cutoff E-value, 0.0001; BLAST coverage, 0.4; number of genomes, 12.

predicting precise specificity of a certain enzyme, the former are intended to attribute proteins to broad functional classes or predict their involvement in the same physiological processes or cellular structures.

Let us consider again the example shown in Figure 2.4. The system of three adjoining SN-cycles links six different enzymes participating in the lysine biosynthesis pathway

(Table 2.1). As seen in Figure 2.4(c), five of these proteins (α, β, γ, δ and ε) catalyze subsequent reactions, while the reaction catalyzed by the enzyme ϕ is separated from the nearest reaction of the first group by two intervening steps, corresponding to a metabolic distance $D = 3$. Assuming normalization coefficient $\lambda_p = 1$, the pathway coefficient (see Chapter 2) will be equal $K_p = 1(5/6) \approx 0.83$ for $D_t = 1$, and $K_p = 1$ for $D_t \geq 3$. Further, all six proteins belong to the same functional role category 01.01.01 (amino acid biosynthesis), which means that the functional category coefficient in this case will be $K_f = 1(6/6) = 1$ (again, assuming $\lambda_f = 1$). Thus, both coefficients indicate a high degree of functional coupling between the enzymes considered. Importantly, none of these three SN-cycles or their parts constitutes a conserved colinear gene cluster, although one such cluster is incidentally present and involves the conserved pair of genes coding for dihydrodipicolinate synthase and homoserine dehydrogenase shared between the *Thermotoga maritima* and *Synechocystis* sp. genomes.

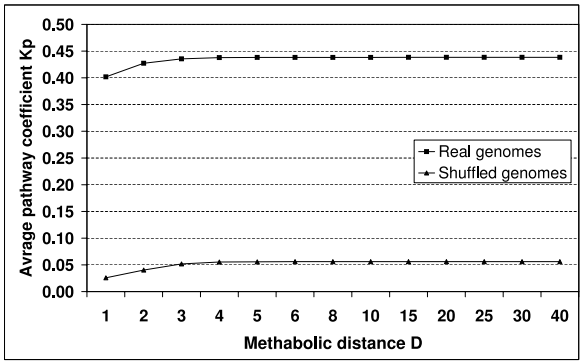
To assess the global performance of our method, we have studied the behavior of the K_p measure on the full set of SN-cycles delineated from 12 genomes. The complete KEGG pathway database (Kanehisa and Goto 2000) was treated as a set of separate subgraphs corresponding to the individual biochemical pathways, such as lysine biosynthesis or glycolysis. Effectively, by using such an approach we are introducing additional *a priori* knowledge about functionally coupled genes in our measurements. Using this approach (Figure 2.6(a)) to estimate K_p leads to a good separation between real and shuffled genomes for all values of the maximally allowed metabolic distance D: the functional content of realistic SN-cycles appears to be an order of magnitude higher. Such bias would not have any influence on K_p if gene groups found by SN-cycles were random.

Comparison of SN-cycles in real and shuffled genomes in terms of the pathway coefficient K_p is presented in Figure 2.6(b). Over 30% of all real SN-cycles found have K_p values greater than 0.5, in contrast to only 1% of random cycles. Even in the range $0.2 < K_p < 0.5$, real SN-cycles have a nearly fivefold lead over the random ones, and the total of 81% of the cycles are in the range $0.2 < Kp < 1.0$. By contrast, the same comparison for the functional category coefficient K_f (Figure 2.6(c)) shows that only 40% of the real SN-cycles are in the range $0.2 < K_f < 1.0$, while 60% have lower K_f values and cannot be statistically distinguished from random cycles. We can thus conclude that the SNAP algorithm is capable of associating gene products involved in a common biochemical pathway, while the specific functions of individual genes represented in terms of a cellular role category appear to be correlated rather weakly.

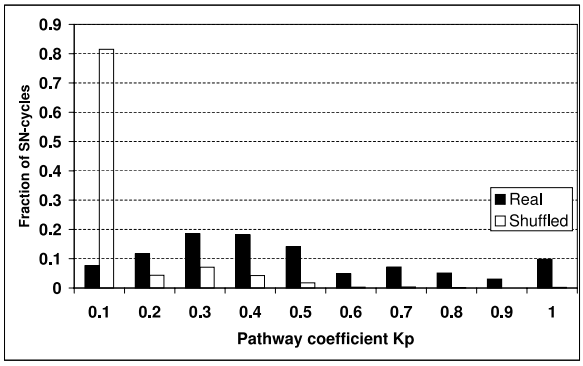
Table 2.1. Genes constituting the SN-cycle shown in Figure 4 (shadowed) and their orthologs.

Genome	Gene	Id	Description	Start	Stop
E.coli	α	g1786214	dihydrodipicolinate reductase	28374	29195
	β	g1790455	lysine-sensitive aspartokinase III	4230812	4229463
	γ	g1788658	usg-1 protein	2434669	2433656
	δ	g1788823	dihydrodipicolinate synthase	2597780	2596902
	ε	g1786183	aspartokinase I/homoserine dehydrogenase	337	2799
	ϕ	g1789203	diaminopimelate decarboxylase	2976921	2975659
C. trachomatis	α	gi_3328787	dihydrodipicolinate reductase	415997	415236
	β	gi_3328785	aspartokinase III	414229	412934
	γ	-	-	-	-
	δ	gi_3328784	dihydrodipicolinate synthase	412923	412063
	ε	-	-	-	-
	ϕ	-	-	-	-
M. tuberculosis	α	rv2773c	dapB dihydrodipicolinate reductase	3082337	3081600
	β	rv3709c	ask aspartokinase	4153480	4152215
	γ	rv3708c	asd aspartate semialdehyde dehydrogenase	4152214	4151177
	δ	rv2753c	dapA dihydrodipicolinate synthase	3067120	3066218
	ε	rv1294	thrA homoserine dehydrogenase	1449373	1450698
	ϕ	rv1293	lysA diaminopimelate decarboxylase	1448026	1449369
T. maritima	α	gi_4982086	dihydrodipicolinate reductase	1516426	1516426
	β	gi_4982084	aspartokinase II	1515057	1513852
	γ	gi_4982089	aspartate-semialdehyde dehydrogenase	1518990	1518007
	δ	gi_4982087	dihydrodipicolinate synthase	1517307	1516423
	ε	gi_4981061	aspartokinase II	574428	572209
	ϕ	gi_4982083	diaminopimelate decarboxylase	1513842	1512682
A pernix	α	-	-	-	-
	β	gi_5104810	473aa long hypothetical aspartate kinase	711805	713226
	γ	gi_5104813	long hypothetical aspartate-semialdehyde dehydrogenase	713223	714272
	δ	-	-	-	-
	ε	gi_5104814	long hypothetical homoserine dehydrogenase	714263	715267
	ϕ	-	-	-	-
Synechocystis sp.	α	gi_1651716	dihydrodipicolinate reductase	77406	77406
	β	gi_1653765	aspartate kinase	3333243	3335045
	γ	gi_1001379	aspartate beta-semialdehyde dehydrogenese	3248483	3249325
	δ	gi_1001380	dihydrodipicolinate synthase	3249385	3250290
	ε	gi_1001182	homoserine dehydrogenase	2627873	2626572
	ϕ	gi_1653772	arginine decarboxylase	3342856	3344943

(a)



(b)



(c)

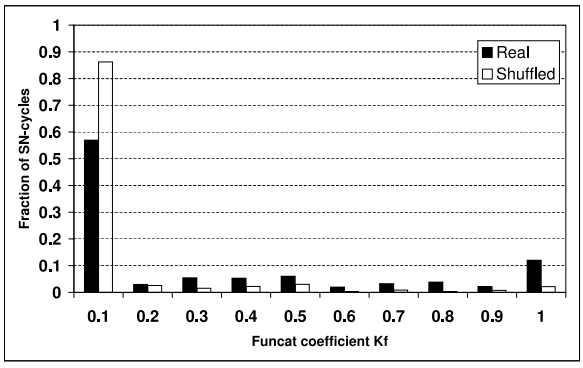


Figure 2.6. Functional content of SN-cycles in real (squares, filled bars) and shuffled (triangles, open bars) genomes. (a) Dependence of the pathway coefficient K_p on the maximal allowed metabolic distance D. (b) Relative occurrence of SN-cycles with different K_p values. (c) Relative occurrence of SN-cycles with different values of the funcat coefficient K_f.

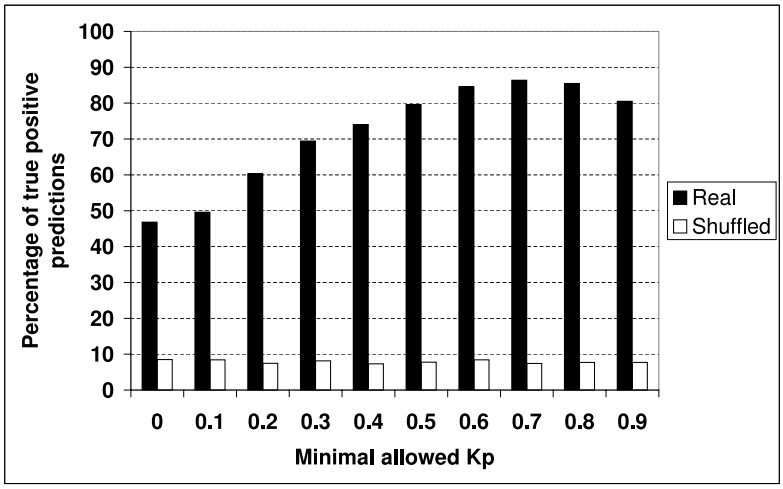


Figure 2.7. Dependence of the percentage of true positive SNAP predictions from the minimal allowed pathway coefficient K_p for real (filled bars) and shuffled (open bars) genomes.

2.8.3. Estimating the predictive power of SNAP

The following simple considerations provide the basis for the estimation of the predictive power of SNAP. Suppose a gene of interest is grouped in an SN-cycle together with a number of other genes with known EC numbers and an arbitrary number of genes without EC numbers assigned. We will ignore the latter, since they make no contribution to the automatic annotation of the query gene. Assuming that at least one gene with a known EC number is related to the query gene, the probability of a correct functional coupling prediction for these particular query gene and SN-cycle is equal to the pathway coefficient K_p of the cycle. However, it may happen that none of the genes in the SN-cycle is pathway-related to the query sequence. Thus, the expected probability of a correct prediction for a given SN-cycle should, on average, be somewhat lower than its K_p, dependent on the frequency of occurrence of a particular functional class. For each gene characterized through SNAP, we calculated K_p of the SN-cycle used for the prediction and compared the pathway assignment of the most represented gene group in the cycle with that of the query gene. Two alternative conditions for considering a prediction of functional coupling to be correct were utilized: (a) best group condition, when the query gene was found in the same pathway as the genes of the single most represented enzyme group in all of the cycles associated with the query gene; and (b) all groups condition, when the query gene was found in the same pathway as the genes of any enzyme group across all cycles.

The cumulative graph in Figure 2.7 shows the dependence of the SNAP best group prediction accuracy on the minimal allowed K_p coefficient based on our data. The average success rate for the entire set of genes participating in the SN-cycle is around 45%. If one considers only SN-cycles with $K_p > 0.4$, the prediction accuracy increases to over 75%. As seen in Figure 2.6(b), approximately 60% of all SN-cycles in real genomes (as opposed to only 7% in shuffled genomes) have the K_p coefficient in this range. Not

surprisingly, the percentage of true positives for the shuffled genomes shown in Figure 2.7 remains constant for all values of K_p. Note that the curve for real SN-cycles in Figure 2.7 tails off somewhat at K_p values greater than 0.9. This happens because many of the SN-cycles with K_p values equal to exactly 1.0 include only two genes with known EC numbers, while SN-cycles with K_p values in the range 0.8-1.0 are typically calculated on the basis of five to ten genes (data not shown). The probability of encountering two out of two genes with the same EC number by chance is higher than, for example, to find eight out of ten genes with the same EC number. In other words, this curve is not normalized by the number of genes actually used to calculate K_p.

Table 2.2. Percentage of true positives for individual genomes and summarized for all genomes.

Genome	Percentage of true positives		Number of genes for which a prediction was made
	Best cycle	All cycles	
A.pernix	78.8	78.8	33
C.jejuni	89.5	91.2	57
C.pneumoniae	84.2	89.5	19
E.coli	72.0	75.2	125
M.pneumoniae	54.5	63.6	11
M.thermoautotrophicum	76.3	76.3	38
M.tuberculosis	85.5	85.5	83
P.abyssi	66.7	76.2	63
Synechocystis. sp	90.3	90.3	62
T.maritima	79.6	79.6	49
T.pallidum	63.6	63.6	11
T.acidophilum	69.0	69.0	58
All genomes	77.8	79.8	609

In Table 2.2 we present the percentage of true positive predictions for the individual genomes studied measured as described above. Only SN-cycles with K_p greater than 0.4 were considered. The best group true positive rate for such cycles varies from 54% for *Mycoplasma pneumoniae* to 90% for *Synechosystis* sp., while the all groups numbers lie in the range from 63% (*M. pneumoniae*, *Treponema pallidum*) to 91% (*Campylobacter jejuni*). Overall, the all groups true positive rate is somewhat better than the best group simply because the odds of finding genes coupled with the query gene in many KEGG pathway maps are higher than in just one map.

2.8.4. Genome annotation with SNAP

The genome of the thermoacidophilic archaeon *Thermoplasma acidophilum* containing 1507 predicted genes has recently been sequenced and subjected to careful manual annotation using the PEDANT software system (Ruepp *et al.* 2000). In particular, each gene was assigned to one of the following categories, reflecting the current level of knowledge about its biochemical function: known protein (24 genes); strong similarity to known protein (189 genes); similarity to known protein (495 genes); weak similarity to known protein (101 genes); strong similarity to unknown protein (110 genes); similarity to unknown

protein (265 genes); weak similarity to unknown protein (85 genes); no similarity (237 genes); and questionable ORF (one gene).

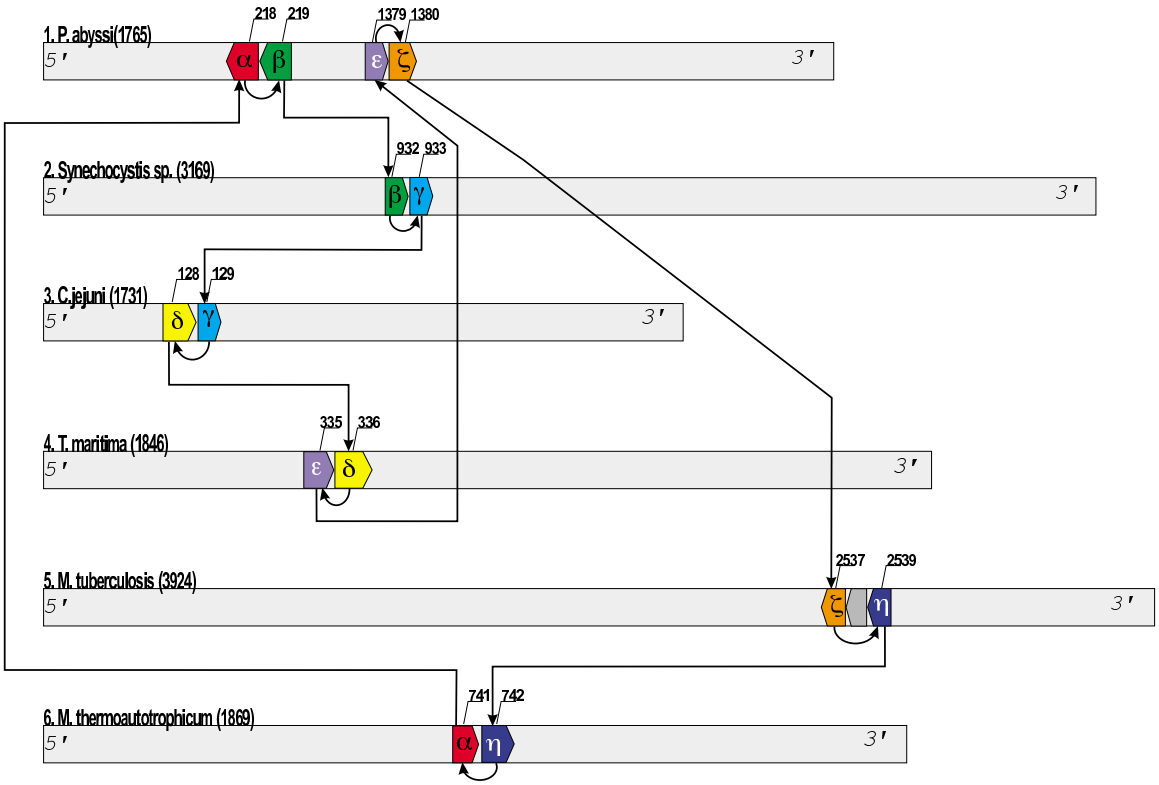
Here, we focus on the 460 *T. acidophilum* genes, or roughly 30% of the gene complement, that possess some degree of similarity to uncharacterized proteins. The number of genes of this type for which a SNAP prediction can be made depends critically on the number of genomes considered and reaches 140, or roughly one-third of this pool, when all 12 genomes are taken into account. This number will definitely grow as more genomes are included in the analysis. It appears that with a sufficient number of phylogenetically distant genomes available, essentially every gene in a genome under scrutiny will participate in at least one SN-cycle.

Let us consider the SNAP results for the *T. acidophilum* gene Ta0740. This gene, described as conserved hypothetical protein, has orthologs in a number of other bacterial genomes, but all of them are functionally uncharacterized. The SN-cycle associated with Ta0740 (denoted α, see Figure 2.8(a)) involves six other types of proteins. Five of them (β, δ, ε, ζ and η) are enzymes with known EC numbers, while the sixth protein, denoted γ, is annotated as chloroplast import-associated channel IAP75. Using our software, we were able to establish that four of the enzymes, δ, ε, ζ and η, catalyze a compact group of biochemical reactions in the phenylalanine, tyrosine, and tryptophan biosynthesis pathway (KEGG map 00400, see Figure 2.8(b)), while the enzyme β and the non-enzymatic protein γ are seemingly unrelated to the first four proteins. Thus, based on these automatically derived KEGG assignments, the value of K_p for this particular SN-cycle is $4/5 = 0.8$, because four out of five proteins with known EC numbers belong to the same metabolic pathway. However, by additional manual analysis we were able to find out that the enzyme b, involved in purine methabolism (KEGG map 00230), is actually only six reactions away from the enzyme ε. Moreover, even the protein γ with no apparent enzymatic activity may be linked to the photosynthesis system that is adjacent to the KEGG map presented in Figure 2.8(b) (see upper left corner). Based on the SNAP results, we predict that Ta0740 is involved in phenylalanine, tyrosine, and tryptophan biosynthesis.

The second example from *T. acidophilum* is a SNAP prediction for the gene Ta0420 (Figure 2.9). In the current annotation, this gene is described as conserved hypothetical protein and has similarity to hypothetical proteins in *Methanobacteirum thermoautotrophicum* and *E. coli*. Based on the comparison with the eukaryotic genome of *Saccharomyces cerevisiae*, functional categories regulation of carbohydrate utilization, other energy generation activities and carbohydrate utilization were assigned automatically by the PEDANT system to this protein; these assignments, however, are based on quite weak similarities and are thus questionable.

SNAP detected two SN-cycles: a short four-node cycle composed of the proteins of α and β types, and a long cycle involving the genes α, β, γ, δ, ε, ζ and η (Figure 2.9(a)). The first cycle represents the case of a weakly conserved colinear gene pair: the genes α and β appear in close proximity in just two relatively close genomes (*M. thermoautotrophicum* and *T. acidophilum*). Consequently, based on the annotation of the gene β, we can putatively assign function to the gene α. Specifically, functional categories automatically assigned to β by PEDANT do indeed coincide with those assigned to α (see above) and

(a)


1. P. abyssi(1765)
5′
3′
α
β
ε
ζ
218
219
1379
1380
2. Synechocystis sp. (3169)
932
933
3. C.jejuni (1731)
128
129
δ
γ
4. T. maritima (1846)
335
336
5. M. tuberculosis (3924)
2537
2539
η
6. M. thermoautotrophicum (1869)
741
742

(b)

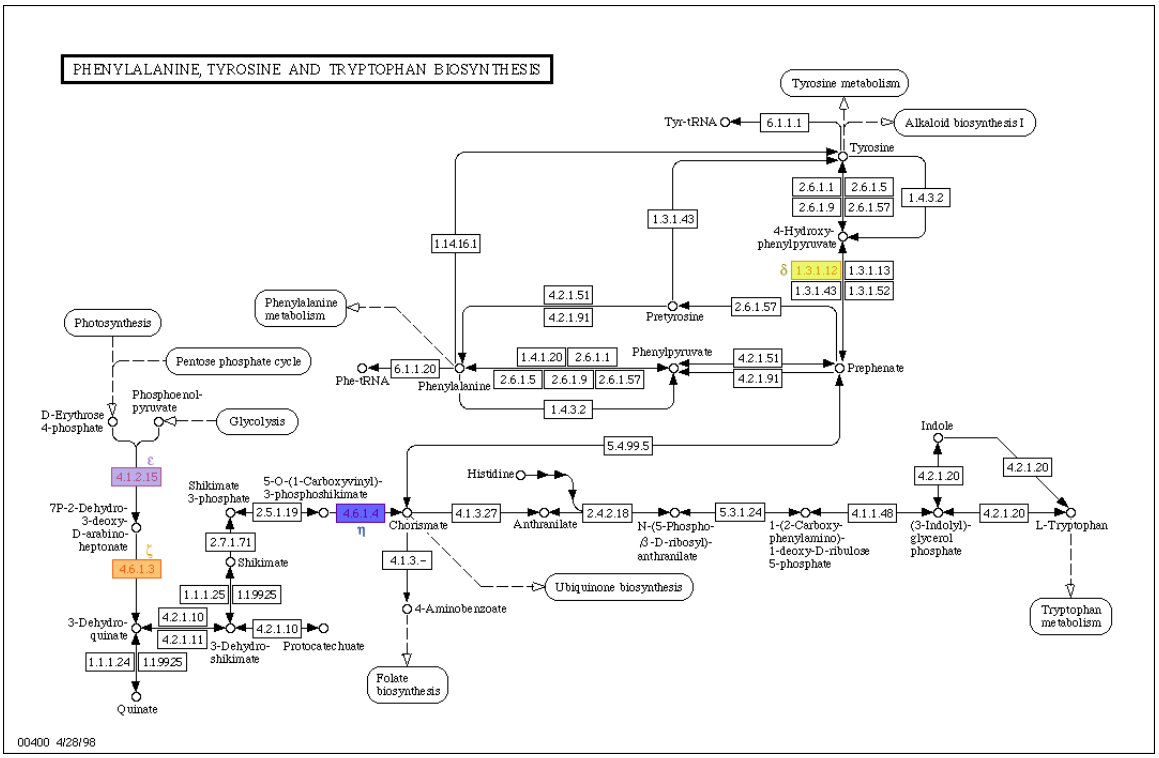
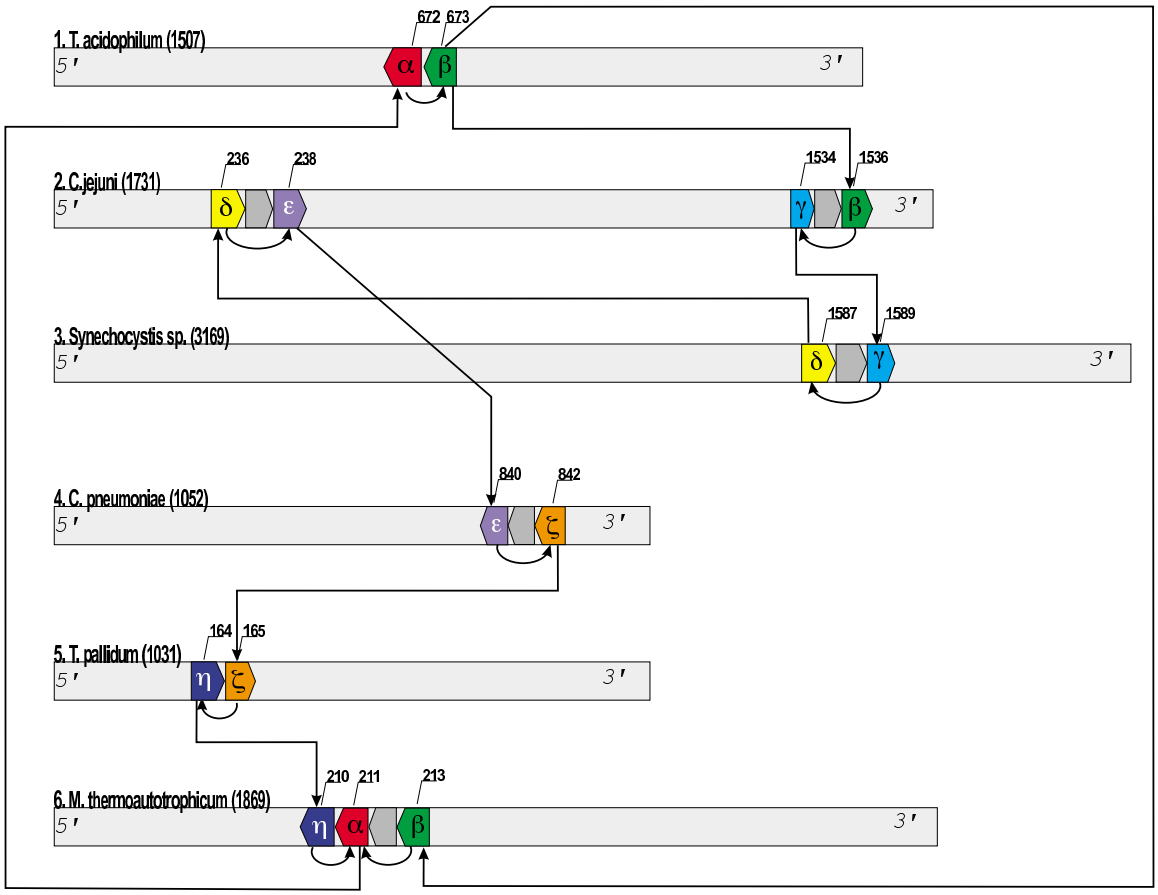

PHENYLALANINE, TYROSINE AND TRYPTOPHAN BIOSYNTHESIS
Tyrosine metabolism
Alkaloid biosynthesis I
Tyr-tRNA
6.1.1.1
Tyrosine
2.6.1.1
2.6.1.5
2.6.1.9
2.6.1.57
1.4.3.2
1.3.1.43
1.14.16.1
4-Hydroxy-phenylpyruvate
1.3.1.12
1.3.1.13
1.3.1.43
1.3.1.52
Phenylalanine metabolism
4.2.1.51
4.2.1.91
Prephenate
Photosynthesis
Pentose phosphate cycle
Phe-tRNA
6.1.1.20
Phenylalanine
1.4.1.20
2.6.1.1
2.6.1.5
2.6.1.9
2.6.1.57
Phenylpyruvate
1.4.3.2
D-Erythrose 4-phosphate
Phosphoenol-pyruvate
Glycolysis
5.4.99.5
Indole
4.1.2.15
Shikimate 3-phosphate
5-O-(1-Carboxyvinyl)-3-phosphoshikimate
2.5.1.19
4.6.1.4
Histidine
Chorismate
4.1.3.27
Anthranilate
2.4.2.18
N-(5-Phospho-β-D-ribosyl)-anthranilate
5.3.1.24
1-(2-Carboxy-phenylamino)-1-deoxy-D-ribulose 5-phosphate
4.1.1.48
(3-Indolyl)-glycerol phosphate
4.2.1.20
L-Tryptophan
7P-2-Dehydro-3-deoxy-D-arabino-heptonate
2.7.1.71
Shikimate
4.6.1.3
4.1.3.-
Ubiquinone biosynthesis
1.1.1.25
1.1.99.25
4-Aminobenzoate
Tryptophan metabolism
3-Dehydro-quinate
4.2.1.10
4.2.1.11
3-Dehydro-shikimate
Protocatechuate
1.1.1.24
Quinate
Folate biosynthesis
00400 4/28/98

Figure 2.8. SNAP analysis of the hypothetical protein Ta0740 from T. acidophilum. (a) SN-cycle associated with Ta0740 (denoted α). Six other protein types found are: β, phosphoribosylaminoimidazolesuccinocarboxyamide synthase (EC 6.3.2.6); γ, chloroplast import-associated channel IAP75; δ, prephenate dehydrogenase (EC 1.3.1.12); ε, 2- dehydro-3-deoxyphosphoheptonate aldolase (EC 4.1.2.15); ζ, 3-dehydroquinate synthase (EC 4.6.1.3); η, chorismate synthase (EC 4.6.1.4). (b) Phenylalanine, tyrosine, and tryptophan biosynthesis pathway as presented in the KEGG database (map 00400). Enzymes δ, ε, ζ, and η are highlighted in colors corresponding to those in (a).

thus confirm them (Figure 2.9(b)).

The long SN-cycle reveals the following: α, β, γ and ζ were assigned to the functional category carbohydrate utilization (β, γ and ζ are well-known enzymes occurring in the glycolysis pathway and other energy-related pathways), gene η is a regulatory protein of unclear function, gene δ is a carbonic anhydrase (whose functional role is also not clear) and gene ε is described as NifU-related protein (Figure 2.9(b)). NifU protein is involved in the nitrogen fixation process in certain soil bacteria and cyanobacteria. In our example, though, it has orthologs in Chlamydia pneumoniae and C. jejuni. The existence of nitrogen fixation genes in these host-dependent prokaryotes would be difficult to explain: it is unlikely that such an organism has the ability to perform energetically expensive atmospheric nitrogen fixation in the presence of already fixed nitrogen, as in the host environment. Thus, we conclude that the description assigned to these proteins based on the weak similarity to the nitrogen fixation genes is incorrect.

(a)


1. T. acidophilum (1507)
672 673
5′
α
β
3′
2. C.jejuni (1731)
236 238
1534 1536
δ
ε
γ
β
3. Synechocystis sp. (3169)
1587 1589
4. C. pneumoniae (1052)
840 842
ζ
5. T. pallidum (1031)
164 165
η
6. M. thermoautotrophicum (1869)
210 211 213

(b)

Gene	Genome	PEDANT ID	Description line	Automatically assigned functional categories	
				Number	description
α	T. acidophilum	Ta0420	conserved hypothetical membrane protein	01.05.04	regulation of carbohydrate utilization
				02.99	other energy generation activities
				01.05.01	carbohydrate utilization
α	M. thermoautotrophicum	gi_2621261	FUN34 related protein	01.05.04	regulation of carbohydrate utilization
				02.99	other energy generation activities
				01.05.01	carbohydrate utilization
β	T. acidophilum	Ta0421	probable acetyl-coenzyme-A synthetase	02.99	other energy generation activities
				01.05.01	carbohydrate utilization
				30.16	mitochondrial organization
				09.01	biogenesis of cell wall
β	C. jejuni	cj1537c	acetyl-CoA synthetase	02.99	other energy generation activities
				01.05.01	carbohydrate utilization
				30.16	mitochondrial organization
				09.01	biogenesis of cell wall
γ	C. jejuni	cj1535c	glucose-6-phosphate isomerase	30.03	organization of cytoplasm
				02.01	glycolysis
				02.04	gluconeogenesis
				01.05.01	carbohydrate utilization
γ	Synechocystis sp.	gi_1653253	glucose-6-phosphate isomerase	30.03	organization of cytoplasm
				02.01	glycolysis
				02.04	gluconeogenesis
				01.05.01	carbohydrate utilization
δ	Synechocystis sp.	gi_1653251	carbonic anhydrase	08.16	extracellular transpor
δ	C. jejuni	cj0237	carbonic anhydrase	08.16	extracellular transpor
ε	C. jejuni	cj0239c	nifU-like protein	01.02.01	nitrogen and sulphur utilization
ε	C. pneumoniae	gi_4377178	NifU-related protein	01.02.01	nitrogen and sulphur utilization
ζ	C. pneumoniae	gi_3322436	phosphoglycerate mutase	01.05.01	carbohydrate utilization
				02.01	glycolysis
				30.03	organization of cytoplasm
ζ	T. pallidum	gi_3322436	phosphoglycerate mutase	01.05.01	carbohydrate utilization
				02.01	glycolysis
				30.03	organization of cytoplasm
η	T. pallidum	gi_3322435	cation-activated repressor protein	-	-
η	M. thermoautotrophicum	gi_2621260	iron dependent repressor	-	-

Figure 2.9. Figure 7. SNAP analysis of the hypothetical protein Ta0420 (α) from T. acidophilum. (a) SN-cycle associated with Ta0740 (denoted α). (b) Functional categories assigned by PEDANT.

Chapter 3

Databases, tools and implementations

In this chapter we will provide description of tools and databases used in and developed for SNAP analysis.

3.9. PEDANT genome database

In our implementation SNAP algorithm builds on the PEDANT genome analysis server (Frishman, Mokrejs *et al.* 2003) which currently contains information on about 180 completely sequenced and unfinished genomes, including large eukaryotic genomes such as *Mus musculus* and *Homo sapiens*.

PEDANT is a versatile genome data access and data analysis system which currently provides automatically pre-computed results of broad range of bionformatics methods, set of tools for cross-genome comparison, quick access via BioRS^{tm} retrieval system, computation and visualization of protein-protein interaction (PPI) networks based on experimental data. PEDANT is based on relational database schema compatible with both MySQL^{tm} and Oracle^{tm} database management systems.

The PEDANT genome set consists of three major sections:

1 Genomes which undergo careful in-depth analysis by the MIPS biologists using the subsystem for manual annotation available in the PEDANT software suite. This section currently includes *Neurospora crassa*, *T. acidophilum*, and *A. thaliana*.

2 Completely sequenced and published genomes. The main source of sequence data for this section, including DNA contigs and ORF nomenclature, is the genomes division of GenBank (Benson *et al.* 2002), although in some cases we obtain data directly from sequencing centers. Whenever possible data manually curated by NCBI staff has been used[14]. If a curated version is not available, original data as submitted by the authors[15] is processed. This section contains 5 eukaryotic, 84 eubacterial, and 16 archaebacterial datasets.

3 Unfinished genomic sequences. Gene prediction is conducted by ORPHEUS (Frishman, Mironov *et al.* 1998) in a completely automatic fashion, usually allowing for large overlaps between ORFs. This leads to many over-predicted ORFs, but ensures that fewer real ORFs are missed. In many cases, the PEDANT database is the only source of annotation for such datasets. This section contains 15 eukaryotic, 51 eubacterial, and 3 archaebacterial datasets.

---

[13] Available at ftp://ftp.ncbi.nih.gov/genomes/Bacteria

[14] Available at ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria

Large volume of false positives is inevitable when exhaustive bioinformatics analysis as performed by PEDANT is conducted. For this reason stringent parameters of bioinformatics methods were used whenever possible. The raw output of the methods is stored in the database to make it available for further examination by human expert.

For each of the roughly 650 000 protein sequences the following pre-computed analyses are available:

(A) Protein function

- BLAST similarity searches against the complete non-redundant protein sequence database.
- Motif searches against the PFAM (Bateman *et al.* 2002), BLOCKS (Henikoff *et al.* 1999), and PROSITE (Falquet *et al.* 2002).
- Predictions of cellular roles and functions is based on the high-stringency BLAST searches against protein sequences which have manually assigned functional categories as defined in the FunCat Functional Catalogue developed by MIPS and Biomax Informatics AG. The FunCat catalogue covers a broad range of biological concepts, including cellular processes, systemic physiology, development and anatomy for prokaryotes and unicellular eukaryotes, plants and animals. In addition, genomes annotated with other vocabularies (such as Gene Ontology) can be mapped to FunCat annotations and thus integrated into the similarity search, as already done for the genomes of *Drosophila melanogaster* and *Caenorhabditis elegans*. At present, we use proteins with manually assigned functional categories of the following species: plant *A. thaliana*, fungi *S. cerevisiae* , eubacterium *Listeria monocytogenes EGD* and archaebacterium *T.acidophilum*. More species-specific catalogues are in preparation and will be available shortly (e.g. bacteria *Bacillus subtilis, Helicobacter pylori, Neurospora crassa*).
- Similarity-based predictions of enzyme nomenclature (EC numbers).
- Similarity-based extraction of keywords and superfamily assignments from the PIR-International sequence database (Barker *et al.* 2000).
- Assignment of sequence to known clusters of orthologous groups, COGs (Tatusov *et al.* 2001).

(B) Protein structure

- Sensitive similarity-based identification of known 3D structures and structural domains. For this purpose, the IMPALA software (Schaffer *et al.* 1999) has been utilized which allows comparison of each gene product with a collection of position specific scoring matrices, or profile library, representing sequences with known three dimensional structure from the PDB database (Berman *et al.* 2000) and sequences of structural domains from the SCOP database (Lo *et al.* 2002). CATH (Pearl *et al.* 2001) domain predictions are being currently added to the database.
- Prediction of transmembrane regions using the TMHMM software (Krogh *et al.* 2001).

- Identification of local low similarity regions and entire non-globular domains based on the SEG algorithm (Wootton and Federhen 1993).
- Prediction of coiled coil motifs (Lupas *et al.* 1991).
- Prediction of protein structural classes (all-a, all-b, a/b).

One of the important aspects of genome analysis involves evalutation of gene duplication and identification of paralogous gene families. In PEDANT this is provided by performing all-against-all PSI-alignment within each genome set. Further, sequences possessing sufficient degree of sequence similarity are joined into single-linkage groups. Additionally, sequences highly similar on domain level, as inferred from sensitive HMMER (Eddy 1998) recognition of PFAM-domains, are also joined into the clusters, even if the corresponding BLAST score is below than preset threshold.

3.10. Genome viewer

Genomic sequences contain vast number of various elements - genetic structures of different levels of organisation, such as chromosomes, genes, exons, introns, operons, promoters, sites of binding of regulatory proteins, splice sites and so on. Each of these elements is positioned in respect to other genome elements, it can often be composed from other simpler genetic structures, and ultimately it is encoded by a nucleotide sequence. Often the information about such elements is not the result of experimental analysis, but rather derived from sophisticated prediction algorithms. As a result of applying different algorithmic approaches many genomic elements are represented by several alternative models.

For manual analysis of the genome it is often necessary to have the compact view of genomic structures of different complexity. Here we describe *Jaba* - genome visualization tool, which has been developed to aide our analysis of prokaryotic genomes and as an annotation tool for *A. thaliana* genome project.

Jaba viewer consists of following panels:

1 A genome panel (Figure 3.1(a)). In this section most of the available genetic structures are displayed. Genetic structures or models located on direct strand of the DNA are displayed above the coordinate ruler, structures located on reverse strand - below. Each type of genetic structure, e.g. GENSCAN gene predictions or annotated genes, is displayed in a separate row. Elements constituting more complex structures, e.g. promoter and exons constituting genes, or BLAST alignment stretches composing one BLAST match, are depicted joined together by the line. If there are several similar BLAST matches available, they are shown coalesced into a single bar, which can be expanded with a mouse click. Additionally, possibility to zoom in and out, print the whole view, filter and reload selected rows on the panel is provided.

2 A sequence panel (Figure 3.1(b)). When the user activates the structure on the genome panel, she may choose to closely inspect the sequence by pressing "Get sequence" button. When the nucleotide sequence is fetched, it is displayed in the main part of the

(a)

(b)

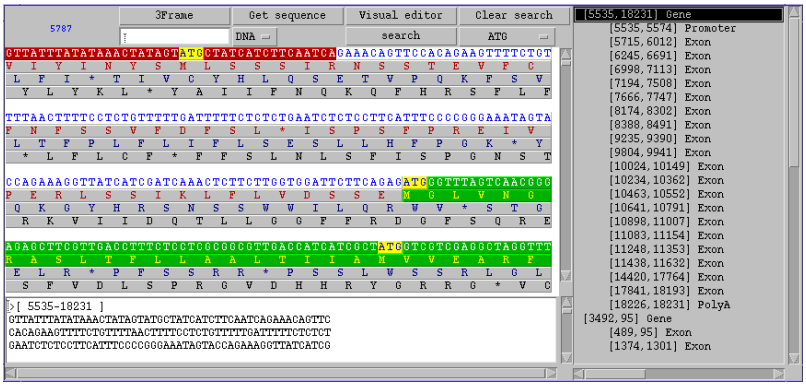


Figure 3.1. Screenshot of *Jaba* genome viewer. (a) Genome browsing panel. Color codes: green - exon, red - promoter, dark cyan - protein BLAST match, lilac - nucleotide BLAST match. Activated gene model is displayed in intensified colors. (b) Sequence and history panels. Color codes are the same as for (a).

sequence panel with substructures (if any) highlighted in the colors corresponding to those on genome panel. Mouse click on the highlighted region results in sequence of the region being pasted into the clipboard panel below. Also, the ability to make three-frame translation of the fetched sequence, find start/stop codons, candidate splice sites is provided.

3 A history panel(Figure 3.1(b)). On this panel the names and coordinates of all previously selected structures and substructures constituting them are displayed in hierarchical manner. Selection of one of the names on this panel activates corresponding structure on the genome panel.

Jaba possesses several features which render it unique among similar programs. First, it is written in *Java* computer language using only the basic *Java* libraries, which makes the viewer immediately accessible via the Internet with any standard browser without requiring user to install any additional libraries or plug-ins. Second, *Jaba* being the full-featured genome viewer is extremely lightweight - executable code fits in only about 60 kilobytes, which makes it ideal for using via the Internet. Third, highly modular design of the viewer allows it to access variety of data sources: SQL database, WWW and flat files. Support of the latter is enormous: *Jaba* supports BLAST, FASTA, GENSCAN, Genemark and numerous other formats, scoring total of about 25 different formats. Furthermore, other formats can be easily added by writing a *Perl* subroutine.

Jaba has being very extensively used in *A. thaliana* genome project (Tabata *et al.* 2000), *Neurospora crassa* genome project and numerous prokaryotic genome projects.

3.11. SNAP Implementation

In the past few years the large number of new prokaryotic genomes have been sequenced. Each new genome is a valuable addition in terms of new context information it contains. However, the size of genomic data itself and the fact that similarity data are essentially binary relations, and being those tend to grow in quadratic fashion, adds up to the complexity of bioinformatics analysis.

SNAP algorithm represents interesting technical challenge in this respect. Although the algorithm is a simple graph traversal, there is no general way of dividing the problem into smaller subproblems. That is, in general SN-graph can not be split into the smaller subgraphs to be traversed on separately, at least not without missing some of the SN-cycles. Therefore, the algorithm must operate on complete SN-graph. That implicates, apart from that the algorithm should run reasonably fast on big and complex graph, it must also have the whole graph in the computer's random access memory (RAM). Fortunately, modern computers are well up to the challenge - it is not generally a problem nowadays to have ~500Mb of RAM required for SN-graph encompassing 25 pro- and eukaryotic genomes. The second problem caused by the dataset size is that initial stage required for fetching the data and building up SN-graph takes considerable time (˜10 min for 25 genomes set) and therefore it is too slow to execute for each single computation.

Having these challenges in mind, we designed our program in the followin way. For performance, SNAP algorithm has been implemented in *C*. It operates in daemon mode, *i.e.* once the data are loaded and the SN-graph is constructed, the program remains resident in the computer's memory accepting connections from external clients and communicating with them by means of a simple text-based protocol. Furthermore, program is able to run several parallel computations on the same SN-graph data structure, thus taking advantage of modern multi-processor architectures. For more extensive computations it has further been parallelized to run on many computers through the use of a network proxy daemon capable of channeling computation requests to the most adequate computer.

3.12. All-against-all alignment

The sensitivity of SNAP algorithm is dependent on the sensitivity of ortholog recognition. The latter might become a crucial problem when evolutionary very distant organisms are present in the dataset, as in case of eukaryotes vs. prokaryotes comparisons (see Chapter 4). For this reason highly sensitive PSI-BLAST (Altschul *et al.* 1997) alignment of 82 protein sets has been conducted.

The amount and the speed with which protein alignment data have been produced turned out to be a hard problem for a standard database management system (DBMS). Consequently, we had to resort to using the time proven flat-file-based approach. In order to retain the ability of running PSI-BLAST computations on several networked computers simultaneously without losing in consistency of computations and the data management, simple database-locking scheme has been implemented using *POSIX* file locking interface. Also, few network daemons has been written to aid alignment data access and management.

After all-against-all alignment has been completed the parsed results have been imported into PEDANT system.

(a)

Gene id	Genome	Description
gi_4981641	T.maritima	adenylosuccinate synthetase
gi_4981642	T.maritima	ornithine carbamoyltransferase, anabolic
g1790620	E.coli	adenylosuccinate synthetase - E. coli
gi_4981640	T.maritima	adenylosuccinate lyase
g1790617	E.coli	hflc protein - E. coli
orf128	T.acidophilum	conserved hypothetical protein
gi_3322377	T.pallidum	Lambda CII stability-governing protein (hflC)
cj0994c	C.jejuni	ornithine carbamoyltransferase
rv1656	M.tuberculosis	argF ornithine carbamoyltransferase
cj0023	C.jejuni	adenylosuccinate lyase
cj0271	C.jejuni	bacterioferritin comigratory protein homolog

(b)

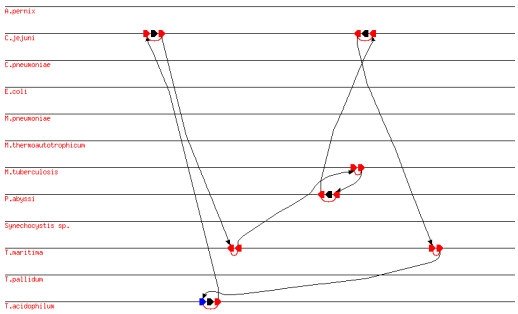


Figure 3.2. SNAP server output for *T. acidophilum* gene coding for adenylosuccinate synthetase. a)The table of genes participating in SN-cycles (sorted by order of occurrence in SN-cycles). The 'Purine metabolism' KEGG/Pathway map is predicted for this gene. b)Diagram of an SN-cycle automatically generated by SNAPper. Blue bar corresponds to gene which starts SN-cycle, red bars to genes contributing to SN-cycle and black bars to genes which lay in-between of two SN-cycle members, but not contribute to SN-cycle themselves.

3.13. SNAPper - SNAP Web server

SNAPper (Kolesov *et al.* 2002) allows to conduct on-line SNAP function predictions for query protein sequences. Using SQL queries, it is possible to correlate pre-computed

properties of gene products stored in PEDANT databases with the results of the SNAPper analysis.

At the time of writing public version of SNAPper utilizes a selection of 23 phylogenetically distant microbial genomes. A SNAPper search can be initiated either with a query protein sequence submitted via a Web form, or by specifying a PEDANT gene id, if known. In the former case a BLAST search against genomic proteins will be performed to find query's orthologues in the genomes considered which will serve as starting points for SNAPping. Subsequently, all SN-cycles originating from the starting points are identified and a list of genes constituting these cycles displayed, equipped with web links to the corresponding PEDANT report pages. This list is expected to be enriched in genes functionally coupled with the query gene. If EC-numbers are available for some part of the genes found, a Web link to the most relevant KEGG (Kanehisa and Goto 2000) metabolic pathway map will be provided. In addition, SNAPper renders a hyperlinked graphical representation of the SN-cycles (Figure 3.2).

The result of the SNAPper analysis does not represent a definitive function prediction, but should rather be considered an aid for further manual annotation. The SNAP analysis parameters (BLAST E-value threshold, criteria for finding orthologues, the number of gene neighbours considered, etc) are set to strict values in order to reduce the number of false positives found, but can be manipulated by advanced users.

3.14. PWP - Phylogenetic Web Profiler

Phylogenetic profiling has become another approach we implemented building upon PEDANT system and all-against-all alignment database. (Wong *et al.* 2003)

Phylogenetic Web Profiler[16] (PWP) is a program which provides the method of phylogenetic profiles over the Internet. It covers set of 72 genomes, including prokaryotic as well as eukaryotic genomes.

The problem of ortholog determination is addressed in this work in several ways: i) by using our database of highly sensitive PSI-BLAST alignments ii) by using as orthologs only best-to-best PSI-BLAST matches by default an iii) by allowing to tweak the parameters of search for each particular case.

Currently, three ortholog determination parameters can be varied by the user.

1 To account for the non-uniform rate of sequence divergence amongst different ortholog families, the option of PSI-BLAST based E-value cutoff is provided. Stringent cutoffs are expected to eliminate false positives in ortholog prediction of more conserved proteins while more relaxed cutoffs will allow detection of more diverged proteins.

2 The option to specify tolerances to differences in length between the query and hit proteins is provided. Comparison of protein lengths will improve ortholog prediction for

[15] Available on-line at http://pedant.gsf.de/pwp.

proteins with conserved domains. However, when orthologs are products of fission events, using stringent length cutoffs may produce false negatives in generated profiles. To help compensate for this phenomenon, the program predicts ortholog fission by searching for adjacent genes coding for non-overlapping regions of the same protein.

3 Finally, the option of comparing annotations between query and hit proteins by word similarity is available as this may help to detect highly sequence divergent orthologs. The annotations are obtained from the PEDANT database (see Chapter 4).

There are other parameters that can be varied to achieve better performance of phylogenetic profiling for particular input. Options allowing to manipulate which NCBI-based evolutionary lineages (Wheeler *et al.* 2000) will be used for profiling are provided. In addition, in order to predict what proteins might act as analogous replacements to the query protein in other organisms (Liberles *et al.* 2002), PWP provides the option of searching for hits that have an inverted profile to that of the query protein.

Chapter 4

SNAPping up eukaryotic genomes

For decades operons or, in general, gene clusters sharing common regulatory region, have been considered to be a feature mostly confined to prokaryotic kingdom.

However, as *Caenorhabditis elegans* genome project continued to carry on, the presence of operon-like gene clusters in this organism has been firmly established. It has been shown that in *C. elegans*, unlike other eukaryotes and similar to prokaryotes, there are groups of genes which are transcribed together by RNA polymerase and (unlike prokaryotes) subsequently cut by special splicing mechanism into isolated mRNAs (Blumenthal and Spieth 1996).

This finding and availability of completely sequenced *C. elegans* genome has prompted us to test SNAP on *C. elegans* and other eukaryotic genomes. As usual, we will start with an example of SN-cycle to demonstrate potential relevance of our approach.

The SN-cycle found by SNAPping *C. elegans* genome and set of prokaryotic genomes is shown on Figure 4.1(a). The SN-cycle starts with gene α in *C. elegans* which has no annotation assigned, but is highly similar (BLAST E-value 1.0E-58) to gene coding for 3-oxoacyl-[acyl-carrier-protein] reductase in *Thermotoga maritima*. The gene residing next to α, β is attributed "propionyl-CoA carboxylase alpha chain precursor" description, however it shares highest degree of sequence similarity with multifunctional acetyl-CoA carboxylase-biotin carboxylase enzyme from *Streptococcus pneumoniae*. Besides, as we were able to find out in ENZYME[17] database (Bairoch 2000) and BRENDA[18] (Schomburg *et al.* 2002) eukaryotic propionyl-CoA carboxylase also possesses biotin-dependent transcarboxylase activity which it shares with enzyme γ, acetyl-CoA carboxylase-carboxyl transferase. Interestingly, the latter enzyme is missing in *C. elegans*, which might imply that enzyme β carries out its function in *C. elegans* or simply lack of proper annotation for this enzyme. At last, gene coding for 3-oxoacyl-[acyl-carrier-protein] synthase, or δ on Figure 4.1, is located in the neighborhood of γ in *Campylobacter jejuni* and in the neighborhood of α in *Aquifex aeolicus*.

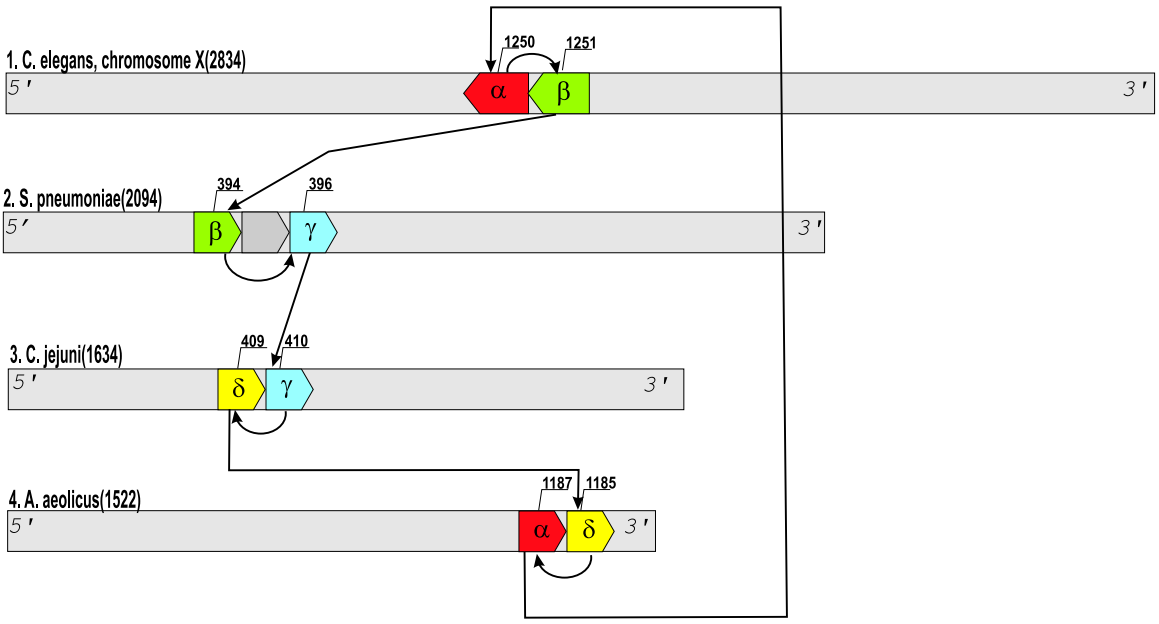
As seen of Figure 4.1(b) the enzymes just described participate in fatty acid biosynthesis pathway. Moreover, these enzymes catalyze the subsequent reactions in the pathway. Thus we can conclude that it is possible to find at least some meaningful SN-cycles connecting together genes in *C. elegans* genome and prokaryotic genomes.

Gene β in *C. elegans* and gene β in *S. pneumoniae*, although highly similar on the sequence level and described as having similar enzymatic activities, are essentially different enzymes. This demonstrates the effect of extending SN-analysis to genomes that are evolutionary very distant from the rest of the genomes considered. In *C. elegans* one of

[16] Available at http://us.expasy.org

[17] Available at http://www.brenda.uni-koeln.de

(a)


1. C. elegans, chromosome X(2834)
1250
1251
α
β
5′
3′
2. S. pneumoniae(2094)
394
396
β
γ
3. C. jejuni(1634)
409
410
δ
γ
4. A. aeolicus(1522)
1187
1185
α
δ

(b)

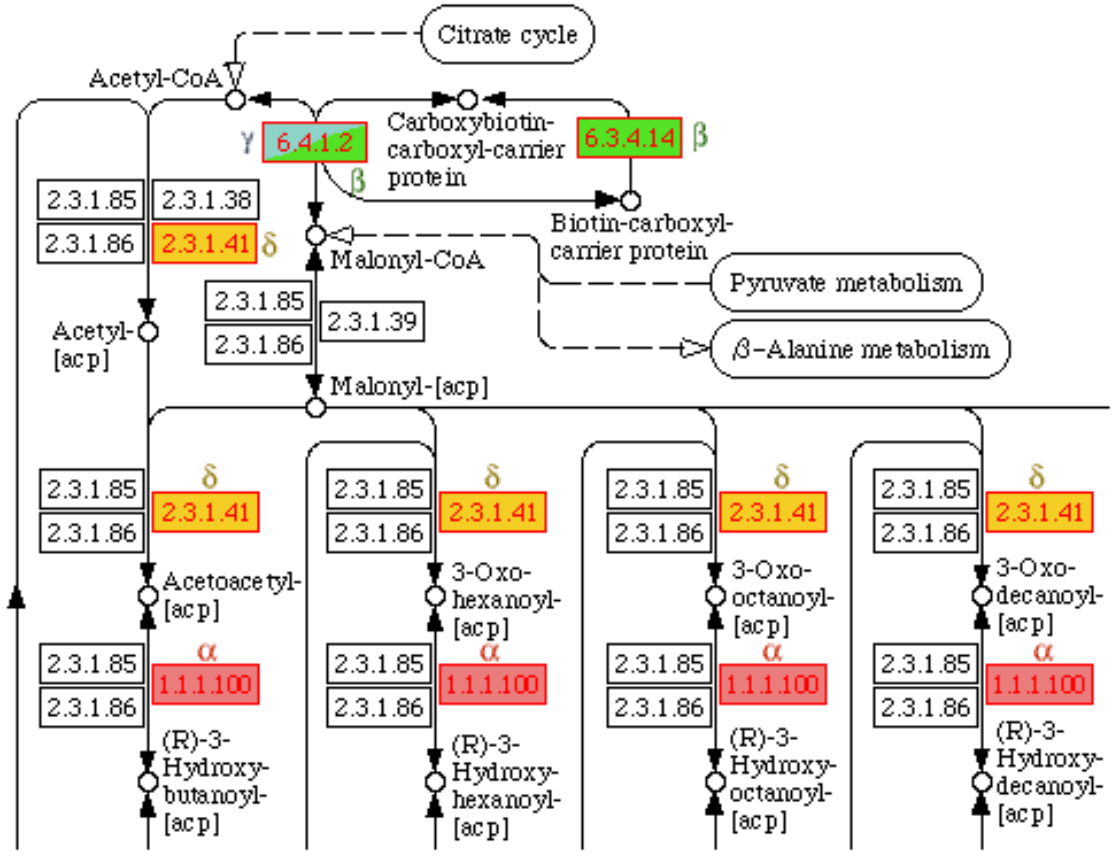

Citrate cycle
Acetyl-CoA
γ
6.4.1.2
Carboxybiotin-carboxyl-carrier protein
6.3.4.14
β
β
2.3.1.85
2.3.1.86
2.3.1.38
2.3.1.41
δ
Malonyl-CoA
Biotin-carboxyl-carrier protein
Pyruvate metabolism
β-Alanine metabolism
2.3.1.85
2.3.1.86
2.3.1.39
Acetyl-[acp]
Malonyl-[acp]
δ
2.3.1.85
2.3.1.86
2.3.1.41
Acetoacetyl-[acp]
3-Oxo-hexanoyl-[acp]
3-Oxo-octanoyl-[acp]
3-Oxo-decanoyl-[acp]
α
2.3.1.85
2.3.1.86
1.1.1.100
(R)-3-Hydroxy-butanoyl-[acp]
(R)-3-Hydroxy-hexanoyl-[acp]
(R)-3-Hydroxy-octanoyl-[acp]
(R)-3-Hydroxy-decanoyl-[acp]

Figure 4.1. (a). A representation of the gene location and their S and N- relationships. The total number of genes in each genome is shown in parentheses. Sequential numbers of genes are indicated. Additionally, each gene is colored and labeled with a Greek letter according to its function: α (red), 3-oxoacyl-[acyl-carrier-protein] reductase (EC 1.1.1.100); β (green), annotated as propyonil-CoA carboxylase in *C. elegans*, highly similar to multifunctional enzyme acetyl-CoA carboxilase-biotin carboxylase in *S. pneumoniae*(EC 6.4.1.3 and 6.4.1.2/6.3.4.14, respectively); γ (light blue) annotated as acetyl-CoA carboxylase - carboxyl transferase (EC 6.4.1.2), δ (yellow), 3-oxoacyl-[acyl-carrier-protein] synthase (EC 2.3.1.41). (b) A part of the KEGG metabolic map involving the four genes predicted to be functionally coupled. Enzymes (highlighted in the same colors as in (a)) encoded by the genes α, β, γ, δ catalyze subsequent reactions in the fatty acid biosynthesis pathway.

the two genes shown codes for propyonil-CoA carboxylase which is involved in fatty acid degradation rather than biosynthesis and is thus not a direct functional analog of the acetyl-CoA carboxylase in *S.pneumoniae* with which it shares the highest sequence similarity (BLAST P-value 1.E-52). As indicated in the ENZYME database an analogous enzyme in plants carries out both functions. This example illustrates the limits of SNAP as a context-based function prediction method: while it is capable of capturing functional relatedness between genes on the coarse level, precise function prediction is usually not possible.

A second example involves genes in the *D. melanogaster* genome. As seen in Figure 4.2 an SN-cycle starts with a pair of reversely oriented genes: α and β. Gene α on D. melanogaster chromosome III codes for the ribosomal protein S10, the product of gene β is described as possessing protein-arginine N-methyltranserase activity, i.e. it is involved in post-translational protein modification. It is homologous to the protein in *Neisseria meningitidis* genome which is annotated as 'conserved hypothetical'. While post-translation modification is in general rare in prokaryotes, the presence of this mechanism has been shown specifically for *N. meningitidis* (Virji 1997; Stimson *et al.* 1996). Although we have not been able to find evidence for protein methylation as a post-translational modification mechanism in prokaryotes this result may hint on its presence in *N. meningitidis*.

The gene γ neighboring gene β is described as "translation elongation factor Tu", also a part of the ribosome, which has its ortholog in *T. maritima*. The neighbor of the latter gene codes for an S10 ribosomal protein which is homologous to the gene α in D. melanogaster described above. All genes in this SN-cycle are involved in translational and post-translational activities of the cell. The gene located between β and γ in the *N. meningitidis* genome codes for a bacterial type ferredoxin. It is not functionally related to other genes considered in this example and does not participate in SN-cycle, demonstrating the property of SN-cycles to avoid such out-of-the-context genes.

4.15. Methods and data

The computational technique used in this work to study gene order in eukaryotic

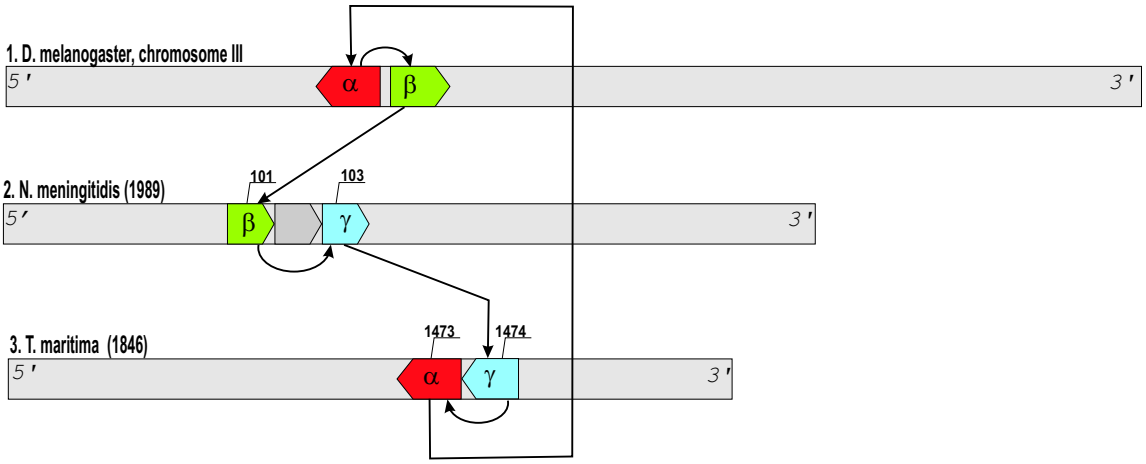


Figure 4.2. Graphical representation of the SN-cycle involving two genes in D. melanogaster genome transcribed in opposite directions. Genes are colored according to their function (red: ribosomal protein S10; green: arginine methyl-transferase/conserved hypothetical protein: blue: translation elongation factor Tu."

organisms is largely based on one described in Chapter 2.

In this study the original SNAP method was modified in several ways to account for larger evolutionary distances and differences in genome structure between prokaryotic and eukaryotic organisms. First, we removed the requirement for the similarity (S-) edges connecting eukaryotic genes to prokaryotic genes or other way around to be the best-to-best PSI-BLAST matches.

Also the S-edges on the SN graph were calculated using less stringent parameters (namely E-value and alignment coverage) for prokaryote-eukaryote comparisons than for prokaryote-prokaryote comparisons. Likewise, while computing the Neighborhood (N) edges on the SN graph the allowed intergenic space size and the number of neighbors on the chromosome considered were different for bacterial and eukaryotic genomes. In addition, in some cases we disregarded all adjacent gene pairs on eukaryotic chromosomes possessing significant sequence similarity and thus representing duplications.

Another difference with respect to the original SNAP method was in the way we retained SN-cycles for analysis. Instead of identifying all SN-cycles having less then a certain number of nodes through the depth-first graph traversal algorithm, as previously described, we performed breadth-first search and selected the shortest SN-cycles containing unique genes not found in any other cycle. This modification allows to prevent combinatorial explosion caused by substantial increase of the number of genomes in our analysis and to decrease the number of spurious SN cycles caused by false positive similarity hits, especially pronounced for comparisons between prokaryotes and eukaryotes. Note that this modification does not cause any changes in the total number of targeted genes participating in SN-cycles. In other words, where the old SNAP scheme (all-SN-cycles, depth-first search) would find some SN-cycles for a particular gene, the new approach

(unique-gene-SN-cycles, breadth-first search) is guaranteed to find a subset of these SN-cycles. As shown on Figure 4.3 this approach changes significantly the distribution of number of SN-cycles of different length.[19]

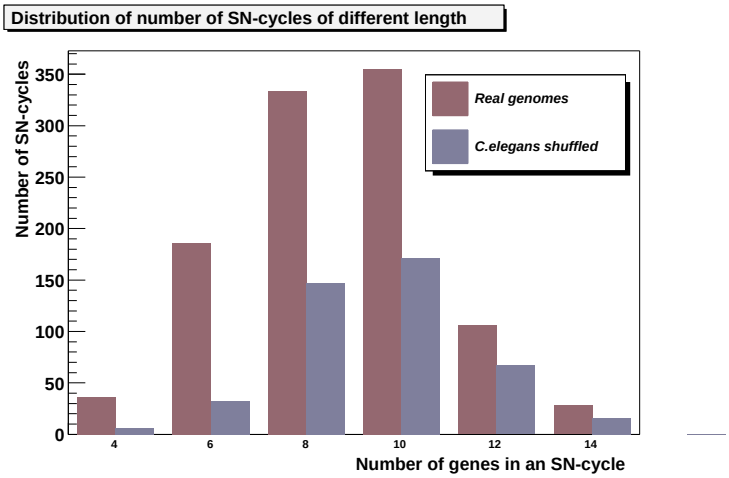


Figure 4.3. Distribution of number of shortest unique-gene SN-cycles of different length originating from genome of worm *C. elegans*.

The same experiment was performed with our set of 23 genomes after randomly shuffling the gene order within studied eukaryotic genome, which effectively leads to destroying meaningful N-relationships in this genome while keeping all S-relationships intact.

Throughout this chapter we will use the signal-to-noise ratio defined as:

$$R = \frac{N_r}{N_s},$$

where N_r is number of pairs of genes in studied genome for which SN-cycles were found and N_s is number of pairs genes found in SN-cycles when studied genome was shuffled. Please note, that in shuffle-test only targeted genome was shuffled, the other (22 prokaryotic genomes) remained unchanged. Also, to account for stochastic oscillations shuffle-test was normally performed several times and the results of the tests were averaged.

4.16. Eukaryotic SN-cycles and the intergenic distance *D*

The size of the intergenic spacer D controls the maximum distance between two genes required for them to be considered involved in an N-relationship, i.e. to be neighbors. In prokaryotic genomes the probability of two genes to be involved in the same operon is intimately connected to D (Overbeek, Fonstein *et al.* 1999). To illustrate the non-random behavior of SN-cycles found in the course of this work in eukaryotic organisms we found

[18] Interestingly, it also provides a hint that our choice of 14-node cut-off on SN-cycle length was optimal.

it convenient to study the dependence of noise ratio *R* between random and shuffled genomes on the intergenic distance *D* using the so-called *DR*-graphs. Throughout this work we present *DR*-graphs both for original gene complements of the genomes studied (dubbed "all") and for the case where tandem duplications between sequence-similar genes were filtered out ("duplication-filtered"). Another critical parameter that had to be taken into account is relative orientation of genes. We distinguished between genes in the head-to-tail orientation having the same direction of transcription ("unidirectional orientation") as well as genes arranged in tail-to-tail ("divergent orientation") and head-to-head ("convergent orientation"). As demonstrated below the performance of the SNAP method on eukaryotic data crucially depends on the organism studied and thus requires organism-specific choice of analysis parameters. For all three eukaryotic genomes considered we could establish a detectable non-random correlation of gene order with prokaryotes under at least a part of parameters tested. In all cases strong influence of tandem duplications was observed.

4.16.1. *DR* dependence in *C. elegans*

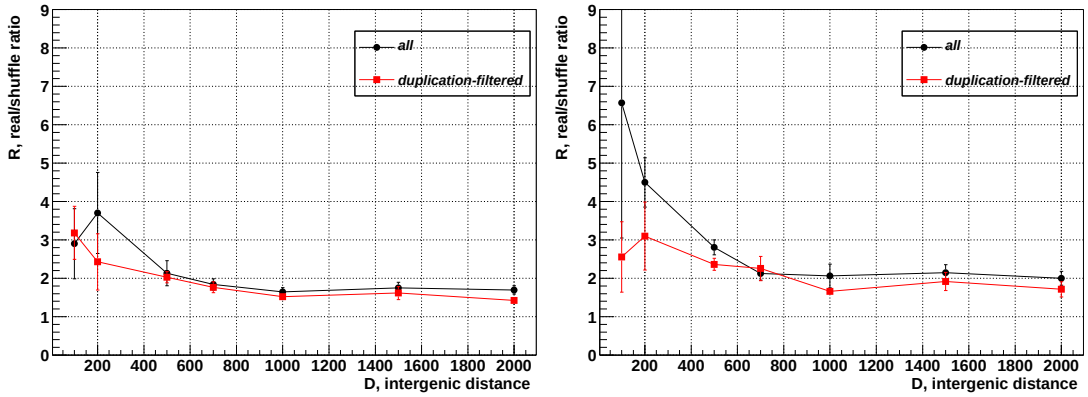
In *C. elegans* signal-to-noise ratio *R* quickly grows with decreasing intergenic spacer size when all gene orientations are considered (Figure 4.4). *R* ratios observed for duplication-filtered SN-cycles are just slightly lower than in the case when all SN-cycles are taken. As in all further tests the behavior of the curves becomes more stochastic towards the lowest values of intergenic distance threshold as the number of genes located in such close neighborhood is very small.

In *C. elegans* SN-cycles involving unidirectional genes and genes in convergent orientation display significant non-randomness ($R>1$) even at large intergenic distances (Figure 4.4(b,d)). Unidirectional genes show similar behavior both with and without tandem duplications as the intergenic distance *D* decreases down to approximately 600 bp. (Figure 4.4(b)). At closer distances the two curves diverge: tandemly duplicated genes appear to make a major contribution to SN-cycles, with *R* values indicating a nearly 7-fold increase over the random level. Duplication-filtered SN-cycles cause a modest albeit quite noticeable increase of *R* in this distance range, implying that unidirectional gene duplications are very prominent in *C. elegans* and usually involve genes located in close proximity to each other.

Surprisingly, genes in convergent orientation display strongly non-random behaviour, although the number of genes in this orientation involved in SN-cycles is relatively small. The *DR* diagram for genes in divergent orientation (Figure 4.4(c)) does not display any significant signal apart from the last data point which is based extremely low number of genes found in such orientation with such strict distance threshold (4 genes) and is thus not statistically significant.

Very strong SN-correlation for unidirectional gene pairs presumably reflects their involvement in operon-like structures The presence of noticeable correlation even at large (> 1000 b.p.) distances may provoke speculations on the presence of a separate

(a) (b)



(c) (d)

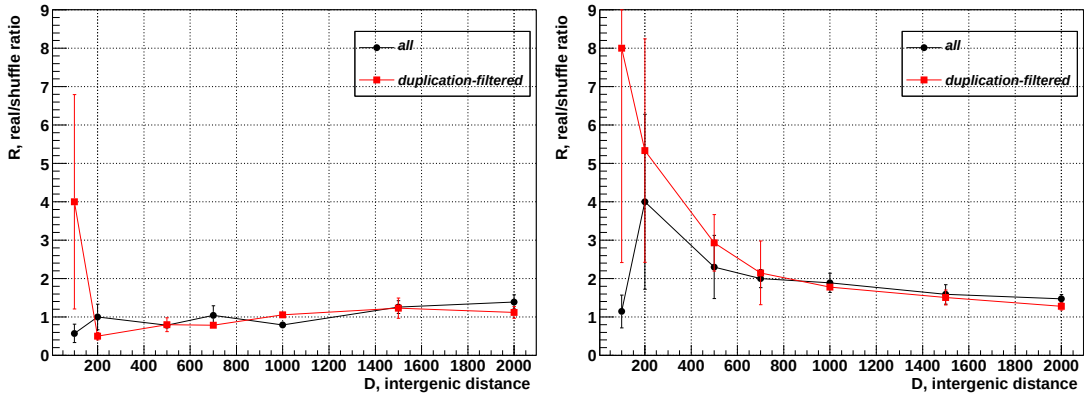


Figure 4.4. *DR*-diagram for *C. elegans* genes participating in SN-cycles broken down for genes in different co-orientation: (a) any orientation (b) unidirectional orientation (c) divergent orientation (d) convergent orientation

mechanism such as chromatin-level regulation which could make these long-distance effects slightly more advantageous in the course of evolution.

Non-random behavior of convergent genes is not easy to explain, since genes in such orientation do not share regulatory zones as divergent gene pairs do, with the only exception of hypothetical long-distance enchancer elements which can be located in-between those genes. Hence, one possibility to explain the result would be to assume existence of enchancer elements located in-between 3' ends of convergent gene pairs regulating transcription of both genes in the pair. We think this is rather unlikely explanation due to various spatial considerations and due to obviously common occurrence of such pairs; also,

other orientations could be affected by such elements as well.

Here we put forward another hypothesis: the products of the genes involved in such convergent gene pairs interact physically directly after translation; translation they undergo is prokaryotic-style translation occurring right after transcription or coupled with transcription. Thus, it could be evolutionary advantageous to place genes coding interacting products in convergent orientation - mRNAs, ribosomes and nascent polypeptides would be closely located aiding genes' products to find each other and interact. The existence of such coupled transcription and translation, and generally translational activity in eukaryotic nucleus has recently been shown in the excellent experimental study by Iborra et al. (Iborra *et al.* 2001)

4.16.2. *DR* dependence in *D. melanogaster*

In *D. melanogaster* (Figure 4.5) non-random correlations between genes in unidirectional and convergent orientation appear to be almost exclusively due to tandemly duplicated pairs, with a dramatic increase of *R* at distances under 3000 base pairs. Unlike *C. elegans* genes in convergent orientation display a weak correlation for $D >$ ~1000 b.p. The graphs for gene pairs with and without duplication filter are virtually indistinguishable.

The only case where we registered significant SN-cycles for duplication-filtered genes was for divergent orientation and small (under 500 b.p.) intergenic distances(Fig. 4.5(c)). A possible explanation for this effect could be sharing of promoter regions between genes in this orientation.

Very strong drop of *R* is observed in duplication-filtered mode for $D<$~200 b.p. for all orientations, but convergent (Fig. 4.5(d)) - the only gene orientation in which two genes do not have promoter region located in-between. That suggests the presence of very strong constraints in *D. melanogaster* genome on minimal intergenic spacer size, most probably on promoter region. The absence of the drop for duplication-unfiltered mode probably implies the presence of large number of recent tandem duplications for which effect of evolutionary pressure on intergenic distance has not yet become detectable. In any event such constraints on the spacer size between duplicated genes are much weaker if present at all.

We examined one of the SN-cycles involving a divergent pair of genes in *D. melanogaster* (see 4.6(a)). This SN-cycle is short, encompassing a gene pair conserved in both *D. melanogaster* and *C. jejuni*. The first gene in this pair codes for a protein described as "ribonuclease III" which is responsible for cleavage and processing of tRNAs, rRNAs, some mRNAs and hnRNAs. The ability to cleave double-stranded DNA has also been demonstrated for this enzyme. The second gene is described as coding for 'ribonuclease HI' and the corresponding enzyme has been shown to possess DNA-RNA hybrid cleavage activity. Evidently, these two enzymes have similar and/or related functions.

On *D. melanogaster* chromosome the start codons of these two genes are separated by a

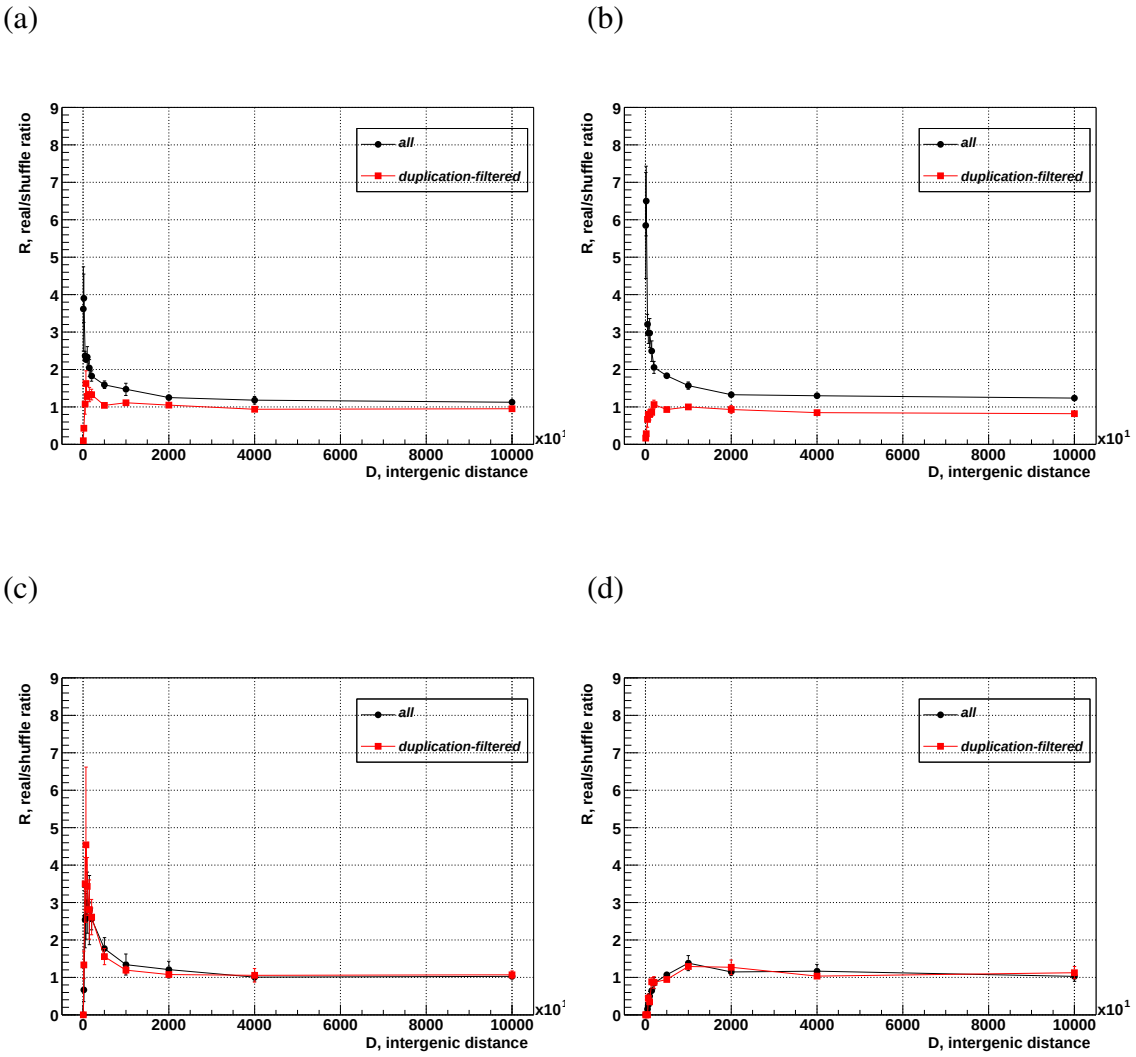
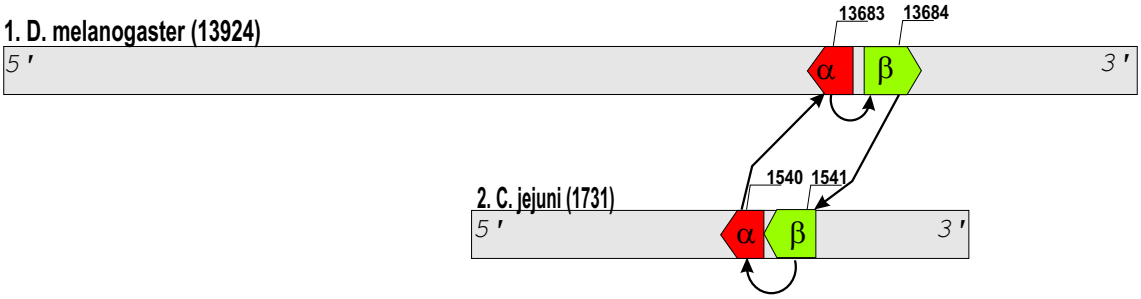


Figure 4.5. *DR*-diagram for *D. melanogaster* genes participating in SN-cycles broken down for genes in different co-orientation: (a) any orientation (b) unidirectional orientation (c) divergent orientation (d) convergent orientation. To account for generally larger intergenic distances in *D. melanogaster* data points up to *D*=100000 b. p. have been computed.

167b.p.-long spacer and the genes are transcribed in opposite directions. To check for candidate regulatory sites within the spacer we extracted the sequence for the spacer and 100 b.p flanking regions at its 5' and 3' ends, and then ran the MatInspector (Quandt *et al.* 1995) transcription factor binding site finding software on obtained sequence. As shown on 4.6(b) the spacer itself, in contrast to the flanking regions, is enriched by putative transcription factor binding sites. The short distances separating these sites and the start codons of both genes suggest that regulatory proteins binding to the sites can affect transcription of both genes, although only wet-lab experiments can provide further evidences supporting this hypothesis.

(a)

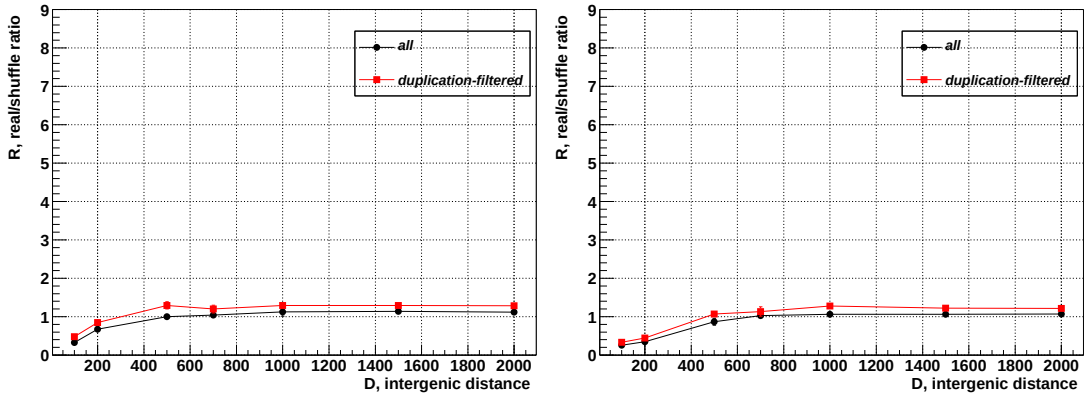


(b)

I$ELF1_01	4	(+)	0.889	0.907	tagtgcgGTGTtgcca
I$BRCZ4_01	73	(-)	1.000	0.883	aataTAAAgcaat
I$CF2II_01	79	(+)	1.000	0.898	ttATATttg
I$DFD_01	101	(+)	1.000	0.934	acgaatATTAggtagt
I$BRCZ4_01	113	(+)	1.000	0.864	tagtTAAAgtaat
I$DFD_01	116	(-)	1.000	0.991	aagttgATTActttaa
I$SN_02	120	(-)	0.906	0.859	ccaagtTGATtact
I$BCD_01	121	(-)	1.000	0.901	ttGATTac
I$SN_01	157	(+)	1.000	0.866	gccacAGGTtcaa
I$ADF1_Q6	168	(+)	0.971	0.850	aaGCTGaagctgcgta
I$BRCZ4_01	193	(-)	1.000	0.930	taatTAAAcgaaa
I$DFD_01	194	(-)	1.000	0.957	agaataATTAaacgaa
I$DFD_01	196	(+)	1.000	0.972	cgtttaATTAttctga
I$E74A_01	217	(+)	1.000	0.881	acaaacgGGAAataaaa
I$BRCZ4_01	221	(+)	0.909	0.893	acggGAAAtaaaa
I$HB_01	225	(+)	1.000	0.865	gaaatAAAAt
I$BRCZ4_01	235	(-)	1.000	0.914	atatTAAAaaaat
I$HB_01	235	(-)	1.000	0.901	ttaaaAAAAt
I$DFD_01	236	(-)	1.000	0.943	tagtatATTAaaaaaa
I$HB_01	236	(-)	1.000	0.874	attaaAAAAa
I$HB_01	237	(-)	1.000	0.922	tattaAAAAa
I$CF2II_01	241	(-)	1.000	0.860	gtATATtaa
I$CROC_01	245	(+)	1.000	0.895	tatacTAAAtaagtta
I$ZESTE_Q2	255	(+)	1.000	0.889	aagttaGAGTgtattg
I$DFD_01	274	(-)	1.000	0.961	agttcaATTAtctcga
I$CROC_01	303	(+)	1.000	0.953	agtcaTAAAtatctcc
I$SN_02	321	(-)	0.906	0.899	agcaccTGCTaaat
I$SN_01	323	(+)	1.000	0.952	ttagcAGGTgctc

Figure 4.6. Example of an SN-cycle involving two genes on *D. melanogaster* chomosome transcribed in opposite directions. (a) Graphical representation of the SN-cycle. Genes are colored according to their function (red: α - ribonuclease III; green: β - riboncuclease HI). (b) MatInspector output showing detected putative regulatory sites in spacer (shaded) separating α and β on *D. melanogaster* chromosome and 100 b.p. regions surrounding spacer. 22 putative sites are located in 167 b.p. spacer in contrast to just 7 in 200 b.p. of flanking regions.

(a) (b)



(c) (d)

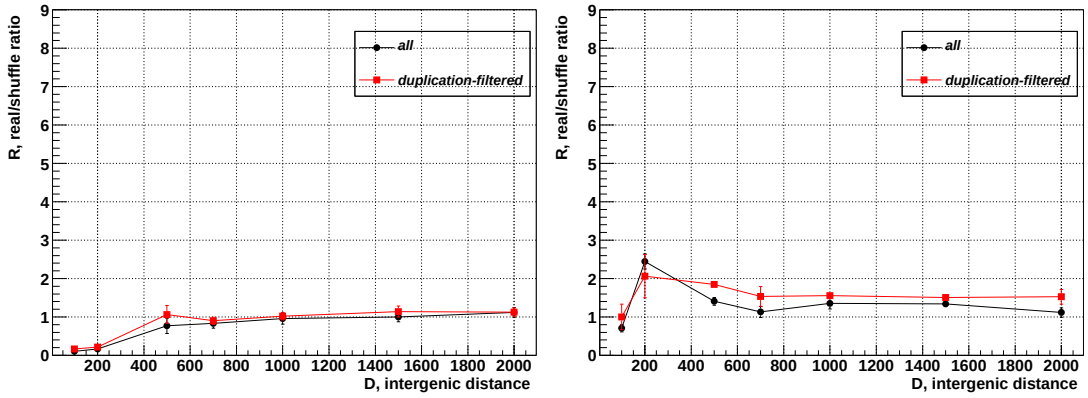


Figure 4.7. *DR*-diagram for *S. cerevisiae* genes participating in SN-cycles broken down for genes in different co-orientation: (a) any orientation (b) unidirectional orientation (c) divergent orientation (d) convergent orientation

4.16.3. *DR* dependence in *S. cerevisiae*

In a recent publication Hurst et al (Hurst *et al.* 2002). have analyzed 166 gene pairs that are co-expressed and conserved in both *S. cerevisiae* and *Candida albicans*. Their findings can be summarized as follows: i) correlation of expression profiles of genes involved in gene pairs increases with decreasing intergenic distance, ii) the proportion of gene pairs conserved in S. cerevisiae and C. albicans grows with decreasing intergenic distance, and iii) among the genes that are highly co-expressed those in divergent orientation are found more frequently than expected from their overall frequency. We would thus

expect to detect similar tendencies with our method - growth of signal-to-noise ratio with decreasing size of intergenic spacer as well as higher R-ratios for divergent gene pairs.

In yeast unidirectional genes perform modestly better than random at large ($D > 600$ b.p.) intergenic distances while at close range R sharply decreases (Figure 4.7(b)). In both cases R values are higher for duplication-filtered genes, implying a small number of local duplications. Gene pairs in convergent orientation (Figure 4.7(d)) display high R values for the distance range 200 - 400 (the hypothesis explaining this phenomenon is the same as for *C. elegans*). Unlike Hurst et al. we were not able to detect any correlation for pairs in divergent orientations, most likely due to the reasons explained below. R for genes in divergent orientation (Figure 4.7(c)) is essentially random for large intergenic distances and becomes worse than randoma (i.e. < 1.0) for adjacent genes.

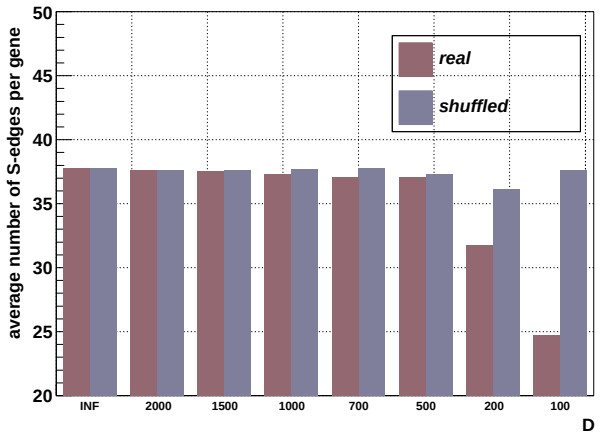


Figure 4.8. Distribution of average number of S-edges per S. cerevisiae gene for different D

In order to find the reason of how algorithm can perform significantly worse than random we computed similarity-ness profile for yeast genes having intergenic spacer on either side (5' or 3') less than D. As shown on Figure 4.8 average number of S-edges connecting yeast genes to genes in prokaryotic genomes (statistic per gene) drops dramatically for $D < 200$ b.p. It explains the behavior of DR-graph: for $D < 200$b.p we find less SN-cycles than in shuffled simply because of significantly lower number of S-edges for the close neighbors in real *S. cerevisiae* genome. But why does S-profile behave this way?

The lower number of S-edges for genes in close neighborhood implies that those genes are, in average, less similar to the genes in other genomes. Since it is difficult to propose a natural process leading to this kind of separation, we assume that this effect is, in fact, artifact of gene prediction. *I.e.* many of the genes occurring in too close neighborhood are incorrectly predicted ORFs or ORFs with incorrectly extended 5'-region - both of these cases would lower the average number of S-edges per gene. To test this hypothesis we computed the fraction of genes having word "questionable" in their description line in dependence on D. As seen on Figure 4.9, the graph grows sharply when reaches

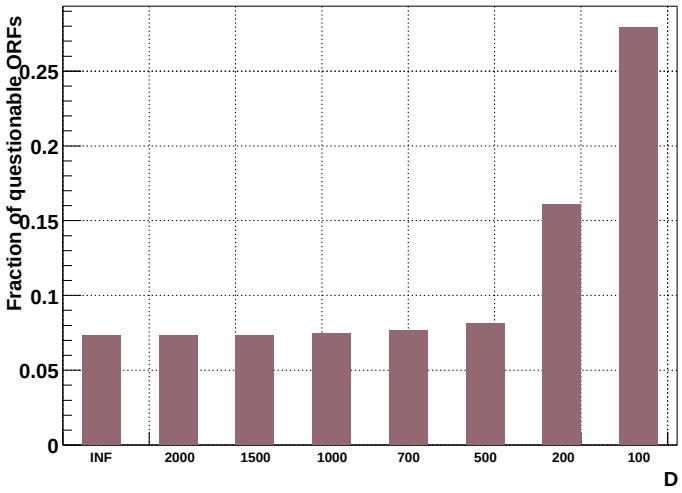


Figure 4.9. Distribution of number of 'questionable' ORFs on D.

$D < 200$, and this growth perfectly corresponds to the drop on the figure 4.8. Thus, the strange behavior of *DR* graph is due to the presence of multiple artifact ORFs sharing an evident tendency to be located in the close neighborhood.

Chapter 5

Conclusion

SNAP is a generalization of the algorithm described by Overbeek et al (Overbeek, Fonstein *et al.* 1998; Overbeek, Fonstein *et al.* 1999). Our method does not rely on the conservation of gene order in the form of colinear gene clusters and detects genes that are functionally coupled through a chain of alternating S and N-relationships. The algorithm takes a protein sequence and a set of annotated completely sequenced genomes as input and returns a number of SN-cycles with all vertices being potentially linked to the query sequence. The main finding that is the wide occurrence of SN-cycles and their strong non-randomness as compared with genomes in which gene order was artificially shuffled. The fact that SN-cycles actually reflect the conservation of gene order makes them a useful instrument for defining functional relationships among genes, studying genome plasticity, and reconstructing evolutionary events. While the biological background of the SN-cycles remains unclear at this point, we assume that they reflect functional coupling between closely co-regulated genes in prokaryotic genomes and, more generally, the conservation of functional and regulatory contexts in genomes (Lathe III *et al.* 2000).

Further, we sought to quantify the ability of SNAP to predict broad gene function. Using assignments of genes to KEGG metabolic maps and the genome annotation available through the PEDANT database, we have demonstrated the tendency of SN-cycles to reveal the proximity of functionally coupled genes. In doing so, our consideration was necessarily limited to the genes to which EC numbers could be assigned. Moreover, the metabolic pathway and functional category assignments that served as the basis for calculating the K_p and K_f coefficients were produced automatically based on sequence similarity searches and are prone to errors. Thus, while the anecdotal evidence of functional coupling detection by SNAP presented throughout this work appears to be quite convincing, objective assessment of SNAP performance is very difficult and is currently limited to recovering rough pathway information for some of the genes involved. Moreover, using this approach we are capable of finding putative true positive predictions, but cannot make any conclusions about negative predictions, i.e. cases when no prediction could be made. In any event, it is clear that the reliability of functional inferences made with SNAP will depend critically on the quality of the whole body of genome annotation available.

Significantly better performance of SNAP in terms of the pathway coefficient K_p as compared with the functional category coefficient K_f is not unexpected and is compatible with the main bulk of facts available on the functional composition of gene clusters. Bacterial operons tend to encode members of distinct protein families required for subsequent steps in a biochemical or regulatory pathway. There is also sufficient evidence that the conservation of spatial proximity is especially pronounced between the physically interacting genes (Itoh *et al.* 1999; Dandekar *et al.* 1998). We have thus confirmed that the concept of functionally coupled or functionally related genes used in context-based prediction methods actually means functionally interacting or jointly acting genes.

We do not claim to provide the algorithmically most optimal approach to exploring SN-relationships in genomes. The filtering criterion for SN-graphs that we used, namely the requirement for SN-paths to be closed, is essentially equivalent to the requirement of two alternative SN-paths between two functionally coupled genes to be present. A more strict criterion would require that more than two alternative paths between two genes exist. We plan to test the performance of SNAP with the number of gene neighbors in each direction considered $c > 2$ (see Chapter 2). Increasing c may allow the detection of long-range patterns in gene order. The main factor limiting the potential of any approach exploiting the conservation of gene order is the massive disruption of gene clusters in distantly related species and the resulting reduction of the number of significant N-relationships available. Another obvious limitation is the possibility of non-orthologous gene displacement (Koonin *et al.* 1996), leading to termination of SN-cycles due to the absence of their constituent S-relationships. The results of the functional coupling prediction are also dependent on our ability to differentiate orthologs of a certain gene in other genomes from paralogous genes. However, even if a homologous protein with a similar function is recruited instead of the true functional ortholog, the SN-graph may still be closed and the corresponding prediction of significant value.

An important recent advance is the establishing of functional association between spatially separated genes that in other organisms are fused to form a composite protein (Marcotte *et al.* 1999; Enright *et al.* 1999). Gene fusion events have been shown to be reliable indicators of protein interaction, but the number of such events is rather limited (e.g. 64 cases involving 2.8 % of proteins in *E. coli*, *Haemophilus influenzae*, and *Methanoccocus jannaschii*, as reported by Enright et al.4). It will be easy to adapt SNAP to take into account gene fusion events by redefining N-relationships as those between separate spatially proximate genes, and those between distinct, non-overlapping sequence domains of the same protein as outlined by the structure of BLAST local alignments. SNAP can also be combined with statistical operon prediction methods (Craven *et al.* 2000) based on recognition of regulatory DNA signals.

Based on our tests with the *Thermoplasma acidophilum* genome, we estimate that SNAP will prove instrumental in mapping functional links for a significant fraction (up to 30 %) of presently uncharacterized genes in bacterial genomes. We plan to launch an effort to re-annotate all completely sequenced genomes available to date. Systematic work directed at the detection of functionally interacting genes will have implications for medical and environmental research, since many genes responsible for antibiotic resistance, pathogenesis, and biodegradation are transferred horizontally between different species in clusters (De La Cruz and Davies 2000) and consequently represent good targets for SNAP. A WWW server allowing the users to perform a gene function prediction using our method and the underlying PEDANT genome database has been implemented.

In the course of our analysis we have developed numerous tools and data repositories. *Jaba* has been created for visual analysis of genomic data and models of genetic structures. SNAPper webserver allows to conduct on-line SNAP analysis for the user input. All-against-all alignment database is a generic resource for extracting groups of orthologous and paralogous genes. Using this database we have developed and made available for public PWP - Phylogenetic Web Profiler, an online system allowing to interactively

apply method of phylogenetic profiling. All-against-all alignment database, PWP and SNAPper have become integral parts of PEDANT genome analysis system.

The role and the frequency of occurrence of gene clusters in eukaryotes is completely open. While operons seem not to be generally present in higher organisms, they do play a significant role in some of them. In the *Caenorhabditis elegans* genome, for example, up to 25% of the genes are organized in polycistronic transcription units (Blumenthal and Spieth 1996). A sizeable number of functionally interacting eukaryotic genes are involved in synexpression groups (Niehrs and Pollet 1999). What part of these genes are physically associated on the chromosome remains unclear. These findings have prompted us to apply our method on eukaryotic genomes.

Although when SNAPping eukaryotic genomes we encountered considerably more noise than in original SNAP study on prokaryotes and we had to modify our method in order to cope with it, non-randomness of SN-cycles and, thus, of gene order in the considered organisms is apparent as shown on *DR* −diagrams.

Perhaps, one of the most surprising results we report here is the presence of significant amount of non-random functionally linked gene pairs in divergent orientation in D. melanogaster. Such structures controlled by region lying in-between have been reported in prokaryotes and *S. cerevisiae* before, but not in higher eukaryotes with their complex genome organization. Similar to the other, simpler, organisms, we suggested that relatively small region in-between divergent genes can control transcription of both genes. Also, our results suggest such corregulated pairs are relatively widespread in *D. melanogaster* genome. The brief study of the region separating a pair of genes predicted by our method to be functionally linked (and they are indeed functionally linked) made our point stronger by demonstrating high concentration of putative transcription factor binding sites in this region.

No less surprising is excess of non-random correlated gene pairs in convergent orientation in *C. elegans* and *S. cerevisiae*. It is very difficult to hypothesize on evolutionary forces driving to this kind of local genomic organization with only exception of the hypothesis that evolution may favor genes in such orientation if their products interact right after the translation **if translation occurs directly on-site**, in the nucleus; indeed, it has recently been reported that about 10% of all translation in mammalian cell takes place in the nucleus (Iborra *et al.* 2001). Thus, from this hypothesis it can be expected that functionally related genes arranged in convergent gene pair undergo prokaryotic-style on-site coupled transcription and translation, with following interaction of the gene products. Obviously, these results as well as this hypothesis require further extensive investigation.

As expected uni-directionally oriented genes in C. elegans appeared to be significantly correlated even after duplication filter has been applied, which is in agreement with the estimation that approximately 25% of genes in *C. elegans* participates in operon-like structures.

In *D. melanogaster* genome for which no operon-like organization has ever been shown, any non-randomness in uni-directional gene order disappears when duplication filter is

applied. The presence of large amount of tandem duplication, and their strong effect on amount of found SN-cycles is observed for all three genomes.

Since SNAP relies on information encoded in the gene order to make functional linkage, it turned out to be sensitive to artificial disruption of the gene order as was demonstrated on *S. cerevisiae* genome. Apparently, introduction of large amount of non-sense ORFs by seeking to fill the gaps gene finder program, obscured almost all non-random correlation of colocalized genes as has been reported by Hurst *et al*. We plan to apply our method again when new release of S. cerevisiae database will come out: as we have been promised, these non-sense ORFs will finally be removed.

We plan to apply our method on eukaryotic organisms from other phyla, in particular plants (*A. thaliana*) and mammals (*Mus. musculus* and *Homo sapiens*).

To draw the line, SNAP method developed by us has been shown as an unique tool capable of uncovering gene function as well as non-trivial gene associations in complex eukaryotic genomes. Moreover, SNAP results strongly hint on the presence of structures and gene arrangements in the eukaryotic genomes which have not been known before. These results can be used to guide further experimental studies. The strength of our method is that it does not rely on any explicit functional information in order to uncover functional association, thus almost entirely avoiding the noise introduced by human errors or by other methods.

References

Altschul *et al.* 1997.

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.,* 25, pp. 3389-3402 (1997).

Bairoch 2000.

A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Res,* 28, pp. 304-305 (2000).

Barker *et al.* 2000.

W.C. Barker, J.S. Garavelli, H. Huang, P.B. McGarvey, B.C. Orcutt, G.Y. Srinivasarao, C. Xiao, L.S. Yeh, R.S. Ledley, J.F. Janda, and *et al.*, "The protein information resource (PIR)," *Nucleic Acids Res,* 28, pp. 41-44 (2000).

Bateman *et al.* 2002.

A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, and E.L. Sonnhammer, "The Pfam protein families database," *Nucleic Acids Res,* 30, p. 276–280. (2002).

Ben-Hur *et al.* 2001.

A. Ben-Hur, D. Horn, H.T. Siegelmann, and V.N. Vapnik, "Support vector clustering," *Journal of Machine Learning Research,* 2, pp. 125-137 (2001).

Benson *et al.* 2002.

D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, and D.L. Wheeler, "GenBank," *Nucleic Acids Res,* 30, pp. 17-20 (2002).

Berman *et al.* 2000.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The Protein Data Bank," *Nucleic Acids Res,* 28, pp. 235-242 (2000).

Blattner *et al.* 1997.

F. R. Blattner, G. III Plunkett, C. A. Bloch, N. T. Burland, and M. Riley *et al.*, "The complete genome sequence of *Escherichia coli* K-12," *Science,* 277, pp. 1453-1462 (1997).

Blumenthal and Spieth 1996.

T. Blumenthal and J. Spieth, "Gene structure and organization in Caenorhabditis elegans," *Curr. Opin. Genet. Dev.,* 6, pp. 692-698. (1996).

Breitkreutz *et al.* 2003a.

BJ. Breitkreutz, C. Stark, and M. Tyers, "Osprey: A Network Visualization System," *Genome Biology,* 4(3), p. 22 (2003a).

Breitkreutz *et al.* 2003b.

BJ. Breitkreutz, C. Stark, and M. Tyers, "The GRID: The General Repository for Interaction Datasets," *Genome Biology,* 4(3), p. R23 (2003b).

Brochier *et al.* 2000.

Celine Brochier, Herve Philippe, and David Moreira, "The evolutionary history of ribosomal protein RpS14," *Trends in Genetics,* 16(28), pp. 529-533 (2000).

Burns *et al.* 1990.

D. M. Burns, V. Horn, J. Paluh, and C. Yanofsky, "Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused Escherichia coli tryptophan synthetase alpha and beta chains.," *J Biol Chem,* 265, pp. 687-706 (1990).

Buxlow 1990.

I. Buxlow, "Preparation of artificial bifunctional enzymes by gene fusion," *Biochem Soc Symp,* 57, pp. 123-133 (1990).

Craven *et al.* 2000.

M. Craven, D. Page, J. Shavlik, J. Bockhorst, and J. Glasner, " A probabilistic learning approach to whole-genome operon prediction," *Intell Syst Mol Biol,* 8, pp. 116-127 (2000).

Dandekar *et al.* 1998.

T. Dandekar, B. Snel, M. Huynen, and P. Brok, "Conservation of gene order: a fingerprint of proteins that physically interact," *Trends Biochem Sci.,* 23, pp. 324-328 (1998).

Daubin *et al.* 2003.

Vincent Daubin, Nancy A. Moran, and Howard Ochman, "Phylogenetics and the Cohesion of Bacterial Genomes," *Science,* 301, pp. 829-831 (2003).

De La Cruz and Davies 2000.

I. De La Cruz and I. Davies, "Horizontal gene transfer and the origin of species: lessons from bacteria," *Trends Microbiol,* 8, pp. 128-133 (2000).

Eddy 1998.

S.R. Eddy, "Profile hidden Markov models," *Bioinformatics,* 14, pp. 755-763 (1998).

Enright *et al.* 1999.

A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events," *Nature,* 402, pp. 86-90 (1999).

Falquet *et al.* 2002.

L. Falquet, M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hofmann, and A. Bairoch, "The PROSITE database, its status in 2002," *Nucleic Acids Res,* 30, pp. 235-238 (2002).

Felsenstein 1989.

J. Felsenstein, "PHYLIP - phylogeny inference package," *Cladistics,* 5, pp. 164-166 (1989).

Frishman, Albermann *et al.* 2000.

D. Frishman, K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, and H. -W. Mewes, "Functional and structural genomics using PEDANT," *Bioinformatics,* 17, pp. 44-57 (2000).

Frishman and Mewes 1997.

D. Frishman and H. W. Mewes, "PEDANTic genome analysis," *Trends Genet,* 13, pp. 415-416 (1997).

Frishman, Mironov *et al.* 1998.

D. Frishman, A. Mironov, H.-W. Mewes, and M. Gelfand, "Combining diverse evidence for gene recognition in completely sequenced bacterial genomes," *Nucl Acids Res* (1998).

Frishman, Mokrejs *et al.* 2003.

D. Frishman, M. Mokrejs, D. Kosykh, G. Kastenmueller, G. Kolesov, I. Zubrzycki, C. Gruber, B. Geier, A. Kaps, K. Albermann, A. Volz, C. Wagner, M. Fellenberg, K. Heumann, and H.-W. Mewes, "The PEDANT genome database," *Nucl Acids Res* (2003).

Harris *et al.* 2003.

T.W. Harris,, R. Lee, E. Schwarz, K. Bradnam, D. Lawson, W. Chen, D Blasier, E Kenny, F Cunningham, R Kishore, J Chan, H-M Muller, A Petcherski, G Thorisson, A Day, T Bieri, A Rogers, C-K Chen, J Spieth, P Sternberg, R Durbin, and L.D. Stein, "WormBase: a cross-species database for comparative genomics," *Nucleic Acids Research,* 31, pp. 133-137 (2003).

Henikoff *et al.* 1999.

S. Henikoff, J.G. Henikoff, and S. Pietrokovski, "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations," *Bioinformatics,* 15, pp. 471-479 (1999).

Hurst *et al.* 2002.

Laurence D Hurst, Elizabeth J. B. Williams, and Csaba Pal, "Natural selection promotes the consrvation of linkage of co-expressed genes," *Trends Genet,* 18(12), pp. 604-606 (2002).

Huynen *et al.* 2000.

M. Huynen, B. Snel, W. Lathe III, and P. Bork, "Predicting protein function by genomic context: quantitative evaluation and qualitative inferences," *Genome Res,* 10, pp. 1204-1210 (2000).

Iborra *et al.* 2001.

Francisco J. Iborra, Dean A. Jackson, and Peter R. Cook, "Coupled transcription and translation within nuclei of mammalian cells.," *Science,* 293, pp. 1139-1142 (2001).

Issel-Tarver *et al.* 2002.

L. Issel-Tarver, K.R. Christie, K. Dolinski, R. Andrada, R. Balakrishnan, C.A. Ball, G. Binkley, S. Dong, S.S. Dwight, D.G. Fisk, M. Harris, M. Schroeder, A. Sethuraman, K. Tse, S. Weng, D. Botstein, and J.M. Cherry, "Saccharomyces Genome Database," *Methods Enzymol,* 350, pp. 329-46 (2002).

Itoh *et al.* 1999.

T. Itoh, K. Takemoto, H. Mori, and T. Gojobori, "Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes," *Mol Biol Evol,* 16, pp. 332-346 (1999).

Jacob *et al.* 1960.

F Jacob, D Perrin, C Sanchez, and J Monod, "[Operon: agroup of genes with expressio coordinated by an operator]," *CR Hebd Seances Acad Sci,* 29(250), pp. 1727-9 (1960).

Kanehisa and Goto 2000.
M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res,* 28, pp. 27-30 (2000).

Kolesov *et al.* 2001.
G. Kolesov, H. -W. Mewes, and D. Frishman, "SNAPping up functionally related genes based on context information: a colinearity-free approach," *J Mol Biol,* 311, pp. 639-656 (2001).

Kolesov *et al.* 2002.
G. Kolesov, H. -W. Mewes, and D. Frishman, "SNAPper: gene order predicts gene function," *Bioinformatics,* 18, pp. 1017-1019 (2002).

Koonin *et al.* 1996.
E. V. Koonin, A. R. Mushegian, and P. Bork, "Non-orthologous gene displacement," *Trends Genet,* 12, pp. 334-336 (1996).

Krogh *et al.* 2001.
A. Krogh, B. Larsson, G. von Heijne, and E.L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *J Mol Biol,* 305, pp. 567-580 (2001).

Lathe III *et al.* 2000.
Warren C. Lathe III, Berend Snell, and Peer Bork, "Gene context conservation of a higher order than operons," *Trends Biochem,* 25, pp. 474-480 (2000).

Lawrence and Roth 1996.
J. G. Lawrence and J. R. Roth, "Selfish operons: horizontal transfer may drive the evolution of gene clusters," *Genetics,* 143, pp. 1843-1860 (1996).

Liberles *et al.* 2002.
D.A. Liberles, A. Thoren, G. von Heijne, and A. Elofsson, "The use of phylogenetic profiles for gene predictions," *Current Genomics,* 3, pp. 131-137 (2002).

Lo *et al.* 2002.
C.L. Lo, S.E. Brenner, T.J. Hubbard,, and A.G. Murzin, "SCOP database in 2002: refinements accommodate structural genomics," *Nucleic Acids Res,* 30, pp. 264-267 (2002).

Lupas *et al.* 1991.
A. Lupas, M. Van Dyke, and J. Stock, "Predicting coiled coils from protein sequences," *Science,* 252, pp. 1162-1164 (1991).

Marcotte *et al.* 1999.
E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg, "A combined algorithm for genome-wide prediction of protein function.," *Nature,* 402, pp. 83-86 (1999).

Mering1 *et al.* 2003.
Christian von Mering1, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel, "STRING: a database of predicted functional associations between proteins," *Nucleic Acids Res,* 31(1), pp. 258-261 (2003).

Mewes *et al.* 1997.
H. W. Mewes, K. Albermann, M. Bahr, D. Frishman, A. Gleissner, J. Hani, K.

Heumann, K. Kleine, A. Maierl, S. G. Oliver, F. Pfeiffer, and A. Zollner, "Overview of the yeast genome," *Nature,* 387, pp. 7-65 (1997).

Mika *et al.* 1999.
S. Mika, G. Rätsch, G. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," *Neural Networks for Signal Processing IX. IEEE,* pp. 41-48 (1999).

Mushegian and Koonin 1996.
A. R. Mushegian and E. V. Koonin, "Gene order is not conserved in bacterial evolution," *Trends Genet,* 12, pp. 289-290 (1996).

Netzer and Hartl 1997.
W. J. Netzer and F. U. Hartl, *Nature,* 388, pp. 343-349 (1997).

Niehrs and Pollet 1999.
C. Niehrs and N. Pollet, "Synexpression groups in eukaryotes," *Nature,* 402, pp. 483-487 (1999).

Omelchenko *et al.* 2003.
M.V. Omelchenko, K.S Makarova, Y.I. Wolf, I.B. Rogozin, and E.V. Koonin, "Evolution of mosaic operons by horizontal gene transfer and gene dispacement *in situ*," *Genome Biology,* 4:R55 (2003).

Overbeek, Fonstein *et al.* 1998.
R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev, "Use of Contiguity on the Chromosome to Predict Functional Coupling," *In Silico Biol,* 1 (1998).

Overbeek, Fonstein *et al.* 1999.
R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev, "The use of gene clusters to infer functional coupling," *Proc Natl Acad Sci U S A,,* 96, pp. 2896-2901 (1999).

Overbeek, Larsen *et al.* 2000.
R. Overbeek, N. Larsen, G. D. Pusch, M. D'Souza, Jr. E. Selkov, N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov, "WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction," *Nucleic Acids Res,* 28, pp. 123-125 (2000).

Pearl *et al.* 2001.
F.M. Pearl, N. Martin, J.E. Bray, D.W. Buchan, A.P. Harrison, D. Lee, G.A. Reeves, A.J. Shepherd, I. Sillitoe, A.E. Todd, and *et al*, "A rapid classification protocol for the CATH Domain Database to support structural genomics," *Nucleic Acids Res,* 29, pp. 223-227 (2001).

Pellegrini *et al.* 1999.
M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proc Natl Acad Sci U S A,* 96, pp. 4285-8. (1999).

Quandt *et al.* 1995.
K Quandt, K Frech, H Karas, E Wingender, and T Werner, "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in

nucleotide sequence data," *Nucleic Acids Res.,* 23, pp. 4878-4884 (1995).

Remm *et al.* 2001.
Maido Remm, Christian E. V. Storm, and Erik L. L. Sonnhammer, "Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons," *J. Mol. Biol,* 314, pp. 1041-1052 (2001).

Ruepp *et al.* 2000.
A. Ruepp, W. Graml, M. L. Santos-Martinez, K. K. Koretke,, C. Volker, and H. W. Mewes, "The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*," *Nature,* 407, pp. 508-513 (2000).

Schaffer *et al.* 1999.
A.A. Schaffer, Y.I. Wolf, C.P. Ponting, E.V. Koonin, L. Aravind, and S.F Altschul, "IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices," *Bioinformatics,* 15, pp. 1000-1011. (1999).

Schölkopf *et al.* 317.
B. Schölkopf, A.J. Smola, and K.-R. Müller, "Kernel principal component analysis," *Advances in Kernel Methods - Support Vector Learning,* MIT Pres (317-352).

Schomburg *et al.* 2002.
I Schomburg, A Chang, and D Schomburg, "BRENDA, enzyme data and metabolic information," *Nucleic Acids Res,* 30(1), pp. 47-9 (2002).

Shimodaira and Hasegawa 1999.
Hidetoshi Shimodaira and Masami Hasegawa, "Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference," *Mol Biol Evol,* 16(8), pp. 1114-2 (1999).

Snel, Bork, and Huynen 1999.
B. Snel, P. Bork, and M. A. Huynen, "Genome phylogeny based on gene content," *Nature genetics,* 21, pp. 108-2 (1999).

Snel, Lehmann *et al.* 2000.
B. Snel, G. Lehmann, P. Bork, and M. A. Huynen, "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene," *Nucleic Acids Res,* 28, pp. 3442-3444 (2000).

Stimson *et al.* 1996.
E Stimson, M Virji, S Barker, M Panico, I Blench, J Saunders, G Payne, E. R. Moxon, A. Dell, and H. R. Morris, "Discovery of a novel protein modification: alpha-glycerophosphate is a substituent of meningococcal pilin," *Biochem. J.,* 316(1), pp. 29-33 (1996).

Tabata *et al.* 2000.
S. Tabata, T Kaneko, Y. Nakamura, H. Kotani, T. Kato, E. Asamizu, N. Miyajima, S. Sasamoto, T. Kimura, T. Hosouchi, K. Kawashima, M. Kohara, M. Matsumoto, A. Matsuno, A. Muraki, S. Nakayama, N. Nakazaki, K. Naruo, S. Okumura, S. Shinpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, S. Sato, M. de la Bastide, E. Huang, L. Spiegel, L. Gnoj, A. O'Shaughnessy, R. Preston, K. Habermann, J. Murray, D. Johnson, T. Rohlfing, J. Nelson, T. Stoneking, K. Pepin, J. Spieth, M. Sekhon, J. Armstrong, M. Becker, E. Belter, H. Cordum, M. Cordes,

L. Courtney, W. Courtney, M. Dante, H. Du, J. Edwards, J. Fryman, B. Haakensen, E. Lamar, P. Latreille, S. Leonard, R. Meyer, E. Mulvaney, P. Ozersky, A. Riley, C. Strowmatt, C. Wagner-McPherson, A. Wollam, M. Yoakum, M. Bell, N. Dedhia, L. Parnell, R. Shah, M. Rodriguez, L.H. See, D. Vil, J. Baker, K. Kirchoff, K. Toth, L. King, A. Bahret, B. Miller, M. Marra, R. Martienssen, W.R. McCombie, R.K. Wilson, G. Murphy, I. Bancroft, G. Volckaert, R. Wambutt, A. Dusterhoft, W. Stiekema, T. Pohl, K.D. Entian, N. Terryn, N. Hartley, E. Bent, S. Johnson, S.A. Langham, B. McCullagh, J. Robben, B. Grymonprez, W. Zimmermann, U. Ramsperger, H. Wedler, K. Balke, E. Wedler, S. Peters, M. van Staveren, W. Dirkse, P. Mooijman, R.K. Lankhorst, T. Weitzenegger, G. Bothe, M. Rose, J. Hauf, S. Berneiser, S. Hempel, M. Feldpausch, S. Lamberth, R. Villarroel, J. Gielen, W. Ardiles, O. Bents, K. Lemcke, G. Kolesov, K. Mayer, S. Rudd, H. Schoof, C. Schueller, P. Zaccaria, H.-W. Mewes, M. Bevan, and P. Fransz, "Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana," *Nature,* 408(6814), pp. 823-6 (2000).

Tatusov *et al.* 2001.

R.L. Tatusov, D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin, "The COG database: new developments in phylogenetic classification of proteins from complete genomes," *Nucleic Acids Res,* 29, pp. 22-28 (2001).

Thanaraj and Argos 1996.

T Thanaraj and P Argos, *Protein Sci.,* 5, pp. 1594-1612 (1996).

The FlyBase Consortium 2003.

The FlyBase Consortium, "The FlyBase database of the Drosophila genome projects and community literature," *Nucleic Acids Research,* 31, pp. 172-175 (2003).

Vapnik 1998.

V.N. Vapnik, *Statistical Learning Theory,* Wiley, New-York (1998).

Vert 2002.

Jean-Philippe Vert, "A Tree kernel to analyze phylogenetic profiles," *Bioinformatics,* 1, pp. 1-9 (2002).

Virji 1997.

M. Virji, "Post-translational modifications of meningococcal pili. Identification of common substituents: glycans and alpha-glycerophosphate - a review," *Gene,* 192, pp. 141-147 (1997).

Wales and Wild 1991.

M. E. Wales and J. R. Wild, "Analysis of structure±function relationships by formation of chimeric enzymes produced by gene fusion," *Methods Enzymol,* 202, pp. 687-706 (1991).

Watanabe *et al.* 1997.

H. Watanabe, H. Mori, T. Itoh, and T. Gojobori, "Genome plasticity as a paradigm of eubac teria evolution.," *J Mol Evol,* 44, pp. 57-64 (1997).

Wheeler *et al.* 2000.

D.L. Wheeler, C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler,

T.A. Tatusova, and B.A. Rapp, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res,* 28, pp. 10-14 (2000).

Wolf *et al.* 2001.

Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin, "Genome alignment, evolution of prokaryotic genome organizatin, and prediction of gene function using genomic context," *Genome Res.,* 11, pp. 356-372 (2001).

Wong *et al.* 2003.

P. Wong, G. Kolesov, D. Frishman, and W. Houry, "Phylogenetic Web Profiler," *Bioinformatics,* 19(6), pp. 782-785 (2003).

Wootton and Federhen 1993.

J.C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Comput Chem,* 17, pp. 149-163 (1993).