Biologically meaningful classification of protein sequences – a bioinformatic approach

der Fakultät für Biologie
der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines Doktors
der Naturwissenschaften

von

Tancred Gilles Frickey
aus New York

vorgelegte
D i s s e r t a t i o n
2005

Tag der mündlichen Prüfung:	20.07.2005
Dekan:	Prof. Dr. Friedrich Schöffl
1. Berichterstatter:	Prof. Dr. Andrei Lupas
2. Berichterstatter:	Prof. Dr. Alfred Nordheim

Table of contents:

Curriculum Vitae

Name:	Tancred Gilles Frickey
Birth:	28.08.74 in New York
Address:	Vöchtingstr. 09
	72076 Tuebingen, Germany
Telephone:	+49 (0)7071 601342 (work)

Father:	Franklin William Frickey (09.07.42)
Mother:	Genevieve Marie Frickey (11.01.43)

Personal:	Unmarried
	US and French nationality
Languages:	
	German (fluent), English (fluent), French (fluent)
Computer skills:	
	Languages: Java, Perl, HTML
	OS: Windows95/98/2000/NT4, Solaris, Linux

Education:

2001-2005	Ph.D. student: Max-Planck Institute for Developmental Biology, Tübingen "Biologically meaningful classification of protein sequences – a bioinformatic approach."
2001	Diploma: Biology, Constance University "Dealing with mutational saturation at the amino acid level"
1996-2001	Biology, Constance University
WS 1995	Cybernetics, Stuttgart University
06.1994	(Baccalaureat) Abitur Kaiserin-Friedrich-Gymnasium Majors: English, Chemistry
1985-94	Kaiserin-Friedrich-Gymnasium
1981-85	Grundschule Ober-Erlenbach

Professional experience:

10.2001-present	Ph.D. student: MPI for Developmental Biology
SS 1997-WS 2001	Student assistant in the Workgroup Meyer Evolutionary Biology, Constance University
08.07.96-04.08.96	Practical at Sanitätshaus Reininger Construction of Prosthetics-, Orthetics
01.06.96-01.07.96	Practical at the architectural design firm Barsties
10.07.95-04.08.95	Practical at ITT Automotive Europe Gmbh (Lathe, milling machine)
04.10.94-10.07.95	National service (France)

Other interests:	Climbing, Swimming, Sculpture, Juggling, Goldsmithying, CAD, Programming.

Publications:

Coles M., Djuranovic S., Soeding J., Frickey T., Koretke K., Truffault V., Martin J., Lupas AN., (2005). AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels. *Structure (Camb)*,**13**(6):919-28.

Proikas-Cezanne T., Waddell S., Gaugel A., Frickey T., Lupas A., Nordheim A., (2004). WIPI-1-alpha (WIPI49), a member of the novel 7-bladed WIPI protein family, is aberrantly expressed in human cancer and is linked to starvation-induced autophagy. *Oncogene,* **23**(58):9314-25.

Frickey, T., Lupas, A.N., (2004) CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**(18): 3702-4.

Santos, L., Frickey, T., Peters, J., Baumeister, W., Lupas, A.N., Zwickl, P. (2004) *Thermoplasma acidophilum* TAA43 is an archaeal member of the eukaryotic meiotic branch of AAA ATPases. *Biol. Chem.* 385(11):1105-11.

Frickey, T., Lupas, A.N., (2004) PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res.*, **32**(17): 5231-8.

Frickey, T., Lupas, A.N., (2004) Phylogenetic analysis of AAA proteins. *J. Struct. Biol.* **146**(1-2): 2-10.

Rabus, R., Ruepp, A., Frickey, T., Rattei, T., Fartmann, B., Starck, M., Bauer, M., Zibat, A., Lombardot, T., Becker, I., Amann, J., Gellner, K., Teeling, H., Leuschner, W.D., Gloeckner, F.-O., Lupas, A.N., Amann, R., Klenk, H.-P. (2004) The genome of Desulfotalea psychrophila, a sulfate-reducing bacterium from permanently cold Arctic sediments. *Environ. Microbiol.* **6**(9): 887-902.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A., Vand de Peer, Y., (2003) Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* **13**(3): 382-90.

Hrbek, T., Kucuk, F., Frickey, T., Stolting, K.N., Wildekamp, R.H., Meyer, A. (2002) Molecular phylogeny and historical biogeography of the Aphanius (Pisces, Cyprinodontiformes) species complex of central Anatolia, Turkey. *Mol. Phylogenet. Evol.* **25**(1):125-37.

Van de Peer, Y., Frickey, T., Taylor, J., Meyer, A. (2002) Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. *Gene* **295**(2):205-11.

Obornik, M., Van de Peer, Y., Hypsa, V., Frickey, T., Slapeta, J.R., Meyer, A., Lukes, J. (2002) Phylogenetic analyses suggest lateral gene transfer from the mitochondrion to the apicoplast. *Gene* **285**(1-2):109-18.

Biologisch relevante Klassifizierung von Proteinsequenzen – ein bioinformatischer Ansatz.

Das Leben wäre ohne Proteine unvorstellbar. Die meisten strukturellen Komponenten des Lebens bestehen aus Proteinen, die meisten metabolischen Reaktionen werden durch Proteine begünstigt und selbst die Vervielfältigung des Erbguts würde ohne Proteine nicht stattfinden. Das Erbgut enthält, in verschlüsselter Form, Informationen über alle Proteine die ein Lebewesen herstellen kann. Will man auf molekularem Niveau Lebewesen verstehen, so ist ein genaues Verständnis der verschiedenen metabolischen und regulatorischen Proteine, sowie deren Interaktionspartner, notwendig. Allerdings ist die experimentelle Beschreibung aller Proteine in allen Organismen sowohl zeitlich als auch finanziell nicht möglich. Um dennoch eine Charakterisierung des Grossteils der Proteine eines Organismus zu ermöglichen macht man sich zunutze, dass verwandte Proteine meist auch ähnliche Struktur und Funktion haben. Ermittelte Charakteristika können somit auf verwandte Proteine übertragen werden. Proteinklassifizierung beschäftigt sich damit, den Verwandtschaftgrad ebenso wie funktionelle und strukturelle Gemeinsamkeiten verschiedener Proteine zu ermitteln.

In dieser Arbeit gehe ich kurz in die Grundlagen der Proteinklassifizierung ein: Sequenzähnlichkeitssuche, Sequenz-alignierung und Stammbaum-erstellung. Die Methoden, ebenso wie ihre Vor- und Nachteile, werden kurz beschrieben und Lösungsansätze für die häufigsten Fehler und Probleme dargelegt.

Die vorgestellten Arbeiten beschreiben zwei unterschiedliche Ansätze zur Klassifizierung von Proteinen, PhyloGenie und CLANS. “PhyloGenie” beschäftigt sich mit der Erstellung und Analyse von Phylomen, der Menge aller Gen-Stammbäume für das jeweilige Proteom eines Organismus. Um abzuschätzen wie gut PhyloGenie im Verhältnis zu alternativen Methoden abschneidet, haben wir zwei Datensätze erneut untersucht: a) Die Menge an lateralem Gen-transfer zwischen Thermoplasma und Sulfolobus (Ruepp et al. 2000) und die Suche nach Genen die die Strahlenflosser spezifische Genomduplikation unterstützen (Taylor et al. 2003). Unsere Analyse des *Thermoplasma acidophilum* Phyloms deutet auf wiederholte Austausche grösserer Bereiche genetischen Materials mit entfernt verwandten Archaebakterien der Familie Sulfolobus hin. Ein Vergleich mit anderen Ansätzen lateralen Gen-transfer aufzudecken zeigt, dass PhyloGenie das vorteilhafteste Verhältnis von Sensitivität zu Spezifität aller untersuchten Methoden erreicht. Eine vergleichende Genomanalyse des unvollständigen *Danio rerio* Genoms zeigt eine weitere Applikation Phylom basierter Analysemethoden. Durch Anwendung von PhyloGenie auf die Fragestellung der Strahlenflosser spezifischen Genomduplikation, konnte die Anzahl an Gruppen orthologer Gene verdoppelt werden, die diese Theorie unterstützen.

Im Gegensatz zu PhyloGenie, welches Organismus-spezifisch arbeitet, behandelt CLANS die Analyse ganzer Proteinfamilien. Eine Proteinfamilie umfasst alle von einem Ur-Protein abstammenden Kopien, die sich im Laufe der Zeit zum Teil stark verändert haben können. Grössere Familien können paraloge und orthologe Untergruppen beinhalten und umfassen oft

mehrere tausend Proteine, wodurch Stammbaumanalysen enorm Zeitaufwendig und schlecht überschaubar werden. Der Ansatz von CLANS beruht auf grafischer Darstellung aller paarweisen Sequenzähnlichkeiten. Dies ermöglicht die Analyse erheblich grösserer Datenmengen und ist unempfindlich gegenüber vielen Problemen der traditionellen Stammbaumerstellung.

Anwendung von CLANS auf die Gruppe der AAA-ATPasen ermöglichte zum ersten Mal eine objektive Beschreibung dieser Familie. Existierende Klassifikationen dieser Familie unterscheiden sich zum Teil erheblich in der Anzahl vorhandener Sequenzen, so dass ein Hauptaspekt dieser Arbeit die Enumerierung aller AAA-ATPasen in der nichtredundanten NCBI Proteindatenbank und Beschreibung der Verwandschatsbeziehungen der einzelnen AAA-subfamilien ist. Die Ergebnisse der AAA-analyse sind biologisch nachvollziehbar und überraschende Vorhersagen, zum Beispiel die Homologie einiger N-Domänen entfernt verwandter AAA-ATPasen, wurden durch zusätzliche Untersuchungen verifiziert.

Die Möglichkeit mit CLANS grosse Mengen an unalignierten Sequenzen zu untersuchen hat dazu geführt, dass es zur Grundlage vieler weiterer Analysen wurde. Als publizierte Beispiele sind hierfür die Analyse des TAA43 Proteins (Santos et al. 2004), eine Beschreibung des Wipi-1-alpha beta-propeller Proteins (Proikas-Czesanne et al. 2004) sowie eine Korrektur der Struktur des AbrB Transkriptionfaktors (Coles et al. in press) anzuführen.

Introduction:

Proteins are the basis of life as we know it. Proteins perform nearly all structural, metabolic, regulatory, catalytic and sensory cellular functions. Although all proteins an organism is capable of producing are encoded in the genome, simple copying of the DNA does not result in a functional cell. Genome sequences provide a basis for understanding the blueprint of an organism, however, only a fraction of the information encoded in the DNA is apparent to us. Most cellular properties emerge from complex interactions between molecules. Proteins interacting with DNA or RNA cause certain genome regions to be transcribed into RNA more frequently than others, the RNA transcripts to be edited via splicing or silencing mechanisms and some to be translated into polypeptide chains. Proteins may then refold these chains, modify them posttranslationally and transport them to specific cellular compartments. These steps, and many more, are necessary to transform the information present at the DNA level into a functional cell.

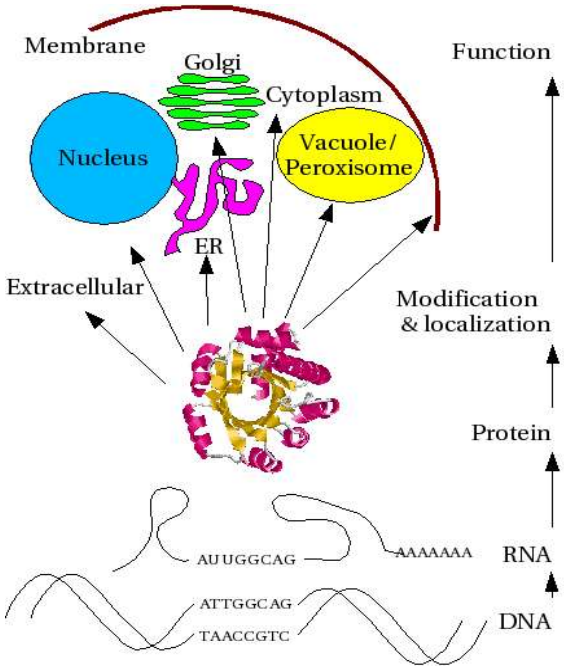


Figure I

Due to the differential expression of genes and subsequent editing and modifying of RNA transcripts and polypeptide chains, the number of protein coding genes is expected to be only a fraction of the number of different proteins present in the organism. The International Human Genome Sequencing Consortium, for example, predicted the estimated 3 billion nucleotides of the human genome to contain 20,000 to 25,000 protein coding genes and be able to generate approximately 1.54 times as many different proteins (IHGSC 2004). The major difficulties on the path to further understanding organism genomes currently lie in determining the proteomes and the protein-protein interactions.

As experimental characterization of all proteins in all genomes would be prohibitively time consuming and expensive, other methods must be used to gather the necessary information about protein structure, function and interaction partners. A widely used approach in genome annotation is to transfer what is known about the closest sequence relatives to the sequences being examined. The relationship of sequences is generally estimated by analyzing their pairwise similarity. Proteins with similar sequence are unlikely to have evolved independently multiple times and therefore must have shared a common ancestor. However, determining closest sequence

relatives is more problematic than it may appear at first glance.

Aligning two sequences:

Dayhoff, Schwartz and Orcutt generated a number of matrices (PAM) in the 1970's that provided estimates of how likely each amino acid was to be replaced by every other. By counting the observed amino acid replacements in alignments of closely related sequences and extrapolating to larger time frames (Dayhoff 1978) they provided a set of matrices describing amino acid substitution probabilities for wide range of evolutionary time. However, their series of matrices were derived from highly similar protein sequences and were shown not to accurately reflect the differences between short-term and long-term substitutions (Gonnet 1992). Many alternative substitution matrices have since been generated, using either larger databases, such as the JTT matrix (Jones 1986), or novel statistical treatment and careful partitioning of the data used to derive the matrices, such as the set of BLOSUM matrices (Henikoff 1991), or the GONNET matrix (Gonnet 1992). Independent of the method used to generate them, all substitution matrices attempt to quantify the probability of each amino acid changing into every other.

```
Ungapped Alignment
       Q E R F G A V P H I R V R K L N A D D G P
       Q E N F G A V V P H L R V K L N A D E G P
Score: 5 5 0 6 6 4 4-2-2-3-2-3-3 5 4 6 4 6 2 6 7
SUM: 55
Gapped Alignment
       Q E R F G A V – P H I R V R K L N A D D G P
       Q E N F G A V V P H L R V – K L N A D E G P
Score: 5 5 0 6 6 4 4-4 7 8 2 5 4-4 5 4 6 4 6 2 6 7
SUM: 88
```

Figure II

The alignment of two sequences is an attempt to place in the same column all residues that evolved from the same position in an ancestral sequence. Adding gaps postulates amino acid insertions or deletions in either of the sequences since divergence from their common ancestor. The information contained in substitution matrices makes it possible to align sequences and insert gaps so as to maximize the statistical probability of the sequences being homologous, i.e. descendant from a common ancestor. The entries in the BLOSUM62 substitution matrix (Appendix A) are logarithms of substitution probabilities. As summing the logarithms is equivalent to multiplying the probabilities, the most likely alignment is the one with the highest value resulting from adding up the substitution matrix entries for all observed amino acid pairings. This value is referred to as the alignment score. The higher the alignment score, the more likely two sequences are to be related. FigureII provides an example of how judicious use of gaps can increase the statistical probability of sequence relatedness. Postulating two insertions, one at position 8 in sequence 2 and one at position 13 in sequence 1, permits alignment of the central region as well as

the N- and C-terminal parts. The score augmentation from aligning the central region is greater than the cost of adding two gaps, resulting in a net increase in alignment score and therefore a higher probability for the two sequences being related.

Finding sequence relatives:

Finding relatives for a given query sequence is generally done via sequence similarity search programs such as BLAST (Altschul 1990), PSIBLAST (Altschul 1997) or FASTA (Lipman 1985). These programs attempt to align a query sequence to all sequences in a database. Estimating the probability of two sequences being related is done by comparing the pairwise alignment score to a distribution of alignment scores for unrelated sequences. The lower the probability of recovering an alignment of score "S" from the distribution, the more likely it is that the sequences are related. The number of false positives, i.e. alignments of unrelated sequences, we expect to find with scores greater than "S", can be calculated as the expect value (E-value) $E=K*m*n*e^{-l*S}$. The parameters "m" and "n" reflect the length of the sequences being aligned, "K" and "l" are scaling parameters that vary with the selected substitution matrix and number of performed sequence comparisons. The first two are used to correct for sequence length, longer alignments tend to have higher scores, while the last two allow results derived from different databases and scoring methods to be compared. NOTE: The E-value is not equal to the probability of two sequences being homologous! This is more closely described by the P-value, the probability of generating at least one alignment with a score equal to or greater than the observed. P-values are calculated as $P=1-e^{-E}$. The smaller the P-value, the more certain one can be in rejecting the null hypothesis, that the alignment score was due to chance similarity of unrelated sequences, and therefore the sequences must be related.

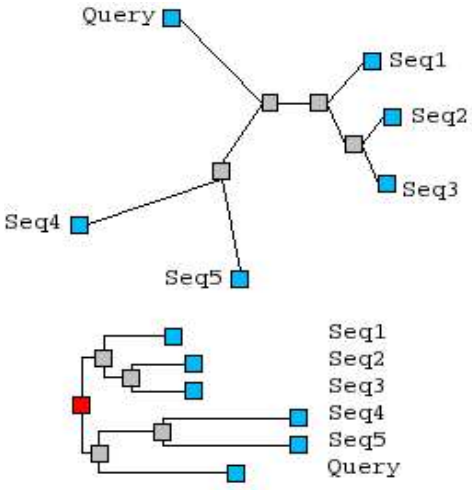


Figure III

During annotation of new, undescribed protein sequences, a frequently used approach is to transfer all that is known about the most similar sequence. This, however, is not the same as transferring the information from the closest sequence relative. Koski & Golding, for example, showed that the best BLAST hit, i.e. the most similar sequence, is not always the closest sequence relative (Koski 2001). A hypothetical example is depicted in FigureIII. The shorter the line connecting two sequences, the more similar they are. Although the query sequence is most similar to sequence1, it cannot be concluded that they are closest relatives. Highest sequence similarity is not necessarily

reciprocal. In this example, sequence1 is more similar to sequence2 and sequence3 than to the query. Calculating all pairwise sequence similarities and inferring a phylogenetic tree resolves the problem and provides a more accurate representation of the evolutionary history. In this case, a midpoint rooting of the tree shows one slowly evolving group with short branch lengths and one more rapidly evolving group with long branch lengths. Based on the best BLAST hit alone, the query sequence would have been classified as a member of the slowly evolving group. The example presented here is only one of the explanations of a well known but frequently overlooked concept, in which the most similar sequence is not identical with the closest sequence relative.

A further example was provided by the IHGSC paper describing completion of the human genome (IHGSC 2001). Based on BLAST results, a vertebrate ancestor was predicted to have acquired 113 genes from bacteria via lateral-gene transfer (LGT). However, subsequent phylogenetic analyses performed by a number of labs, were unable to support the LGT claim for any of the genes in question (Salzberg 2001, Roelofs 2001, Stanhope 2001). This example highlights the importance of using phylogenetic reconstructions instead of highest sequence similarity to determine sequence relatedness.

Phylogenetic tree construction:

Phylogenetic reconstruction methods are predominantly based on multiple sequence alignments. Just as pairwise sequence alignments attempt to place related residues of two sequences in the same alignment column, multiple sequence alignments attempt to place related positions of many sequences in the same alignment column. The information present in the multiple alignment makes it possible to recreate the most plausible path of evolution for this set of sequences, i.e. to infer a phylogenetic tree.

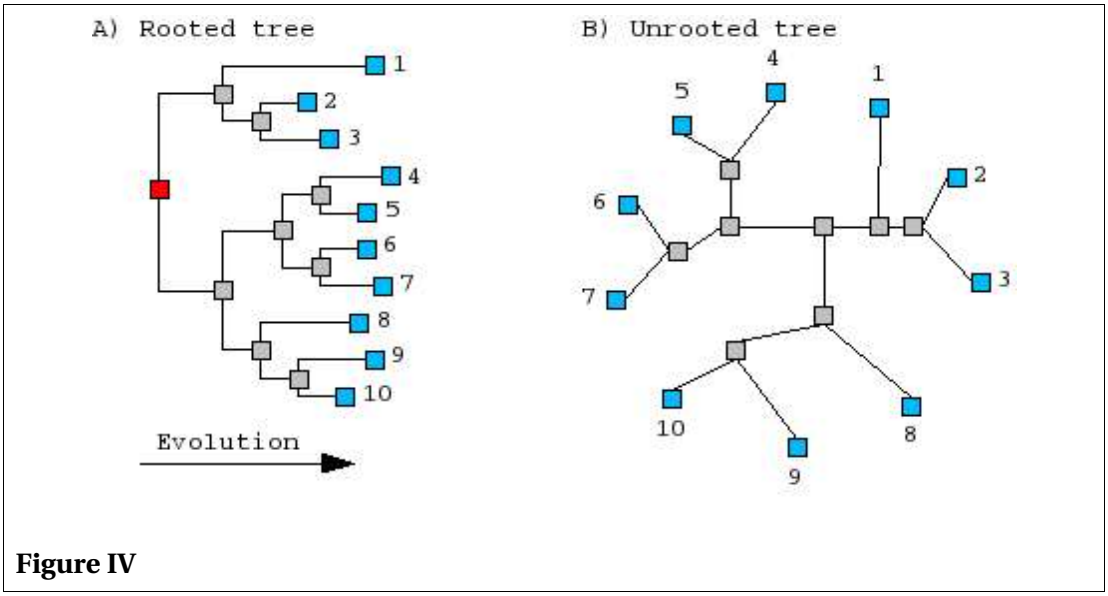


Figure IV

Phylogenetic trees are the visualization of an evolutionary scenario for a group of sequences. They exist in two forms: rooted and unrooted

(FigureIV). While unrooted trees provide only an overview of how closely each sequence is related to every other, rooted trees contain additional information about the direction of evolution, such as which were the ancestral sequences, where did the evolution of this family begin and how much have the individual sequences changed since. The tip nodes, also referred to as operational taxonomic units (OTU's) or leaves of a tree (blue squares), symbolize the sequences present in the alignment. The internal nodes, or branch points (gray squares), symbolize the common ancestor of all descendant OTU's. The rooted tree has a root node (red square) that is used to place the last common ancestor of all sequences in the tree.

Phylogenetic inference programs can be divided into three major groups: maximum parsimony, maximum likelihood and distance methods (FigureV). Maximum parsimony tries to find, for the given alignment, the path of evolution requiring the least number of amino acid changes. Maximum likelihood attempts to reconstruct the most likely path of evolution and distance methods use pairwise comparisons of all sequences to infer a tree. Examples for these three approaches are shown in FigureVI. The examples provide only coarse approximations of the methods and are used only to highlight their differences.

Theoretically, both maximum likelihood and maximum parsimony methods (FigureVI A & B) need to examine all possible trees before being able to determine the best. The number of different unrooted trees for a given set of "N" sequences is: (3)x(5)x...x(2N-5). For four sequences, it is possible to construct three different unrooted trees, for five, there are 15 possibilities, for ten, 2,027,025 and for 15, 7.9×10^{12}. The number of trees that need to be analyzed makes these methods slow and cumbersome for large alignments as their attempts to infer ancestral sequences for every branching point in every possible tree is computationally costly.

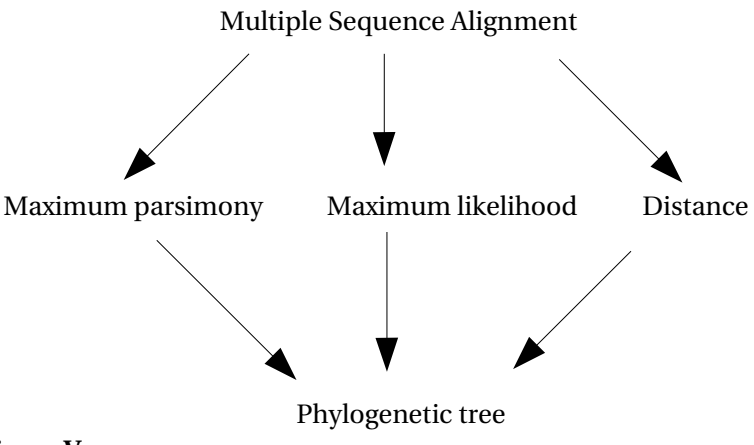


Figure V

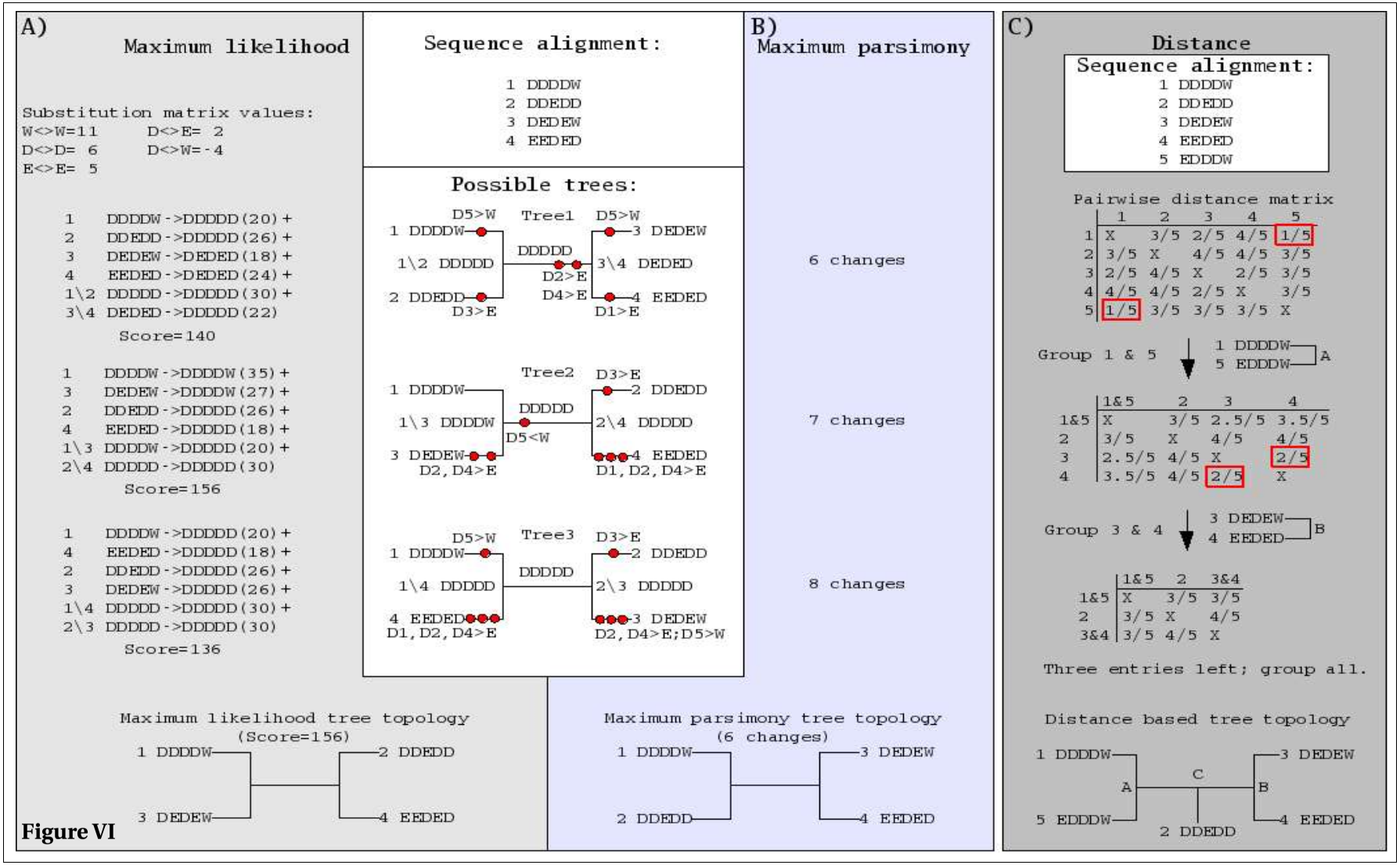

A)
Maximum likelihood
Substitution matrix values:
W<>W=11 D<>E= 2
D<>D= 6 D<>W=-4
E<>E= 5
1 DDDDW->DDDDD(20)+
2 DDEDD->DDDDD(26)+
3 DEDEW->DEDED(18)+
4 EEDED->DEDED(24)+
1\2 DDDDD->DDDDD(30)+
3\4 DEDED->DDDDD(22)
Score=140
1 DDDDW->DDDDW(35)+
3 DEDEW->DDDDW(27)+
2 DDEDD->DDDDD(26)+
4 EEDED->DDDDD(18)+
1\3 DDDDW->DDDDD(20)+
2\4 DDDDD->DDDDD(30)
Score=156
1 DDDDW->DDDDD(20)+
4 EEDED->DDDDD(18)+
2 DDEDD->DDDDD(26)+
3 DEDEW->DDDDD(26)+
1\4 DDDDD->DDDDD(30)+
2\3 DDDDD->DDDDD(30)
Score=136
Sequence alignment:
1 DDDDW
2 DDEDD
3 DEDEW
4 EEDED
Possible trees:
Tree1
D5>W
1 DDDDW
1\2 DDDDD
2 DDEDD
D3>E
DDDDD
D2>E
D4>E
3 DEDEW
3\4 DEDED
4 EEDED
D1>E
Tree2
D3>E
1 DDDDW
1\3 DDDDW
3 DEDEW
D2,D4>E
DDDDD
D5<W
2 DDEDD
2\4 DDDDD
4 EEDED
D1,D2,D4>E
Tree3
D5>W
1 DDDDW
1\4 DDDDD
4 EEDED
D1,D2,D4>E
DDDDD
D3>E
2 DDEDD
2\3 DDDDD
3 DEDEW
D2,D4>E;D5>W
B)
Maximum parsimony
6 changes
7 changes
8 changes
Maximum likelihood tree topology
(Score=156)
1 DDDDW
2 DDEDD
3 DEDEW
4 EEDED
Maximum parsimony tree topology
(6 changes)
1 DDDDW
3 DEDEW
2 DDEDD
4 EEDED
C)
Distance
Sequence alignment:
1 DDDDW
2 DDEDD
3 DEDEW
4 EEDED
5 EDDDW
Pairwise distance matrix
1 2 3 4 5
1 X 3/5 2/5 4/5 1/5
2 3/5 X 4/5 4/5 3/5
3 2/5 4/5 X 2/5 3/5
4 4/5 4/5 2/5 X 3/5
5 1/5 3/5 3/5 3/5 X
Group 1 & 5
1 DDDDW
5 EDDDW
A
1&5 2 3 4
1&5 X 3/5 2.5/5 3.5/5
2 3/5 X 4/5 4/5
3 2.5/5 4/5 X 2/5
4 3.5/5 4/5 2/5 X
Group 3 & 4
3 DEDEW
4 EEDED
B
1&5 2 3&4
1&5 X 3/5 3/5
2 3/5 X 4/5
3&4 3/5 4/5 X
Three entries left; group all.
Distance based tree topology
1 DDDDW
5 EDDDW
A
C
2 DDEDD
B
3 DEDEW
4 EEDED

Figure VI

Maximum parsimony (MP) counts the minimal number of necessary mutations in each tree; Tree1: (1) DDDDW to the ancestral DDDDD=1 change, (2) DDEDD to DDDDD=1 change, (3) DEDEW to DEDED=1 change, (4) EEDED to DEDED=1 change, (1&2 ancestor) DDDDD to DDDDD=0 changes and (3&4 ancestor) DEDED to DDDDD=2 changes. In tree1 the global ancestral sequence may have been DEDED instead of DDDDD, and this is taken into account by the methods when calculating larger trees, but in this example the omission of the alternative makes no difference. Comparable to MP methods, maximum likelihood (ML) calculates the probability of all changes required by the tree. Once all possible trees have been analyzed, either the most parsimonious or most likely, depending on the method, is returned. ML or MP approaches have certain advantages over distance based methods in that the statistics underlying tree inference are well understood and a scenario of sequence evolution is modeled and tested for plausibility. However, even though a number of computational shortcuts have been developed to reduce the time needed for tree calculation, such as Branch-and-Bound or Puzzling (Hendy 1982, Felsenstein 1981, Strimmer 1996), both approaches are still prohibitively time consuming for larger alignments.

An alternative is provided by distance based methods. These do not try to explicitly reconstruct the ancestral sequences and derive an evolutionary scenario, but instead group sequences together based on pairwise distances. Sequence groups and pairwise distances can be generated in a multitude of manners (Sokal 1958, Saitou 1987, Studier 1988, Van de Peer 2002), but grouping is mostly performed by combining pairs of sequences and distance values are generally based on sequence dissimilarity. For simplicity, the example (FigureVI-C) treats the number of differences for each sequence pair as their distance and sequences are grouped together in pairs. To generate the tree, entries with the smallest pairwise distance are selected and combined in a node. In this case, entries '1' and '5' are selected, their distance is 1/5, and combined in node 'A'. The corresponding rows and columns are removed from the matrix and averaged to provide the new distance values for the putative ancestral sequence 'A'. Then the next pair is selected, combined and the matrix updated. This is repeated until only three entries are left. As the final step in tree inference, these three are grouped together. The main advantage of distance methods is their speed. As they do not need to sample all possible trees but only analyze one matrix of size N by N, they are very fast. The statistical basis for tree inference is not well understood and no evolutionary scenario is tested during tree construction, but distance methods generally produce reasonable results and, due to their speed advantage, are able to quickly analyze large datasets. In addition, since these methods are computationally cheap, statistical sampling approaches that rely on comparison of many trees can be used to provide a confidence estimate for every node.

Bootstrapping:

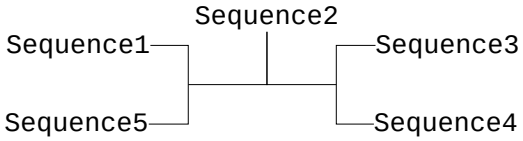
Original alignment

```
Column     12345
Sequence1  DDDDW
Sequence2  DDEDD
Sequence3  DEDEW
Sequence4  EEDED
Sequence5  EDDDW
```

Matrix

	1	2	3	4	5
1	x	2	2	4	1
2	2	x	4	4	3
3	2	4	x	2	3
4	4	4	2	x	3
5	1	3	3	3	x

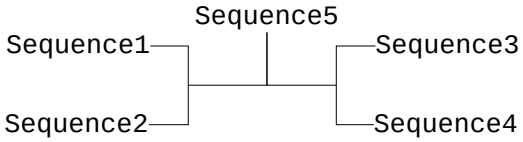
Original tree (distance)



Replicate 1; random values:1,1,5,5,3

```
Column     11553
Sequence1  DDWWD
Sequence2  DDDDE
Sequence3  DDWWD
Sequence4  EEDDD
Sequence5  EEWWD
```

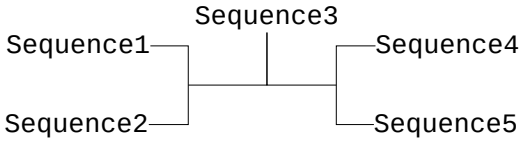
	1	2	3	4	5
1	x	3	0	4	2
2	3	x	3	3	5
3	0	3	x	4	2
4	4	3	4	x	2
5	2	5	2	2	x



Replicate 2; random values:1,5,4,4,1

```
Column     15441
Sequence1  DWDDD
Sequence2  DDDDD
Sequence3  DWEED
Sequence4  EDEEE
Sequence5  EWDDE
```

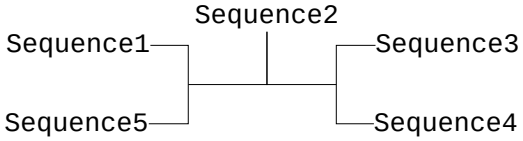
	1	2	3	4	5
1	x	1	2	5	2
2	1	x	3	4	3
3	2	3	x	3	4
4	5	4	3	x	3
5	2	3	4	3	x



Replicate 3; random values:3,2,4,3,3

```
Column     32433
Sequence1  DDDDD
Sequence2  EDDEE
Sequence3  DEEDD
Sequence4  DEEDD
Sequence5  DDDDD
```

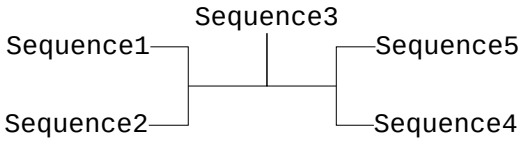
	1	2	3	4	5
1	x	3	2	2	0
2	3	x	5	5	3
3	2	5	x	0	2
4	2	5	0	x	2
5	0	3	2	2	x



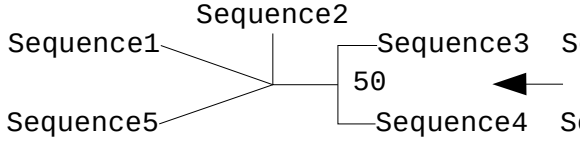
Replicate 4; random values:1,4,2,1,1

```
Column     14211
Sequence1  DDDDD
Sequence2  DDDDD
Sequence3  DEEDD
Sequence4  EEEEE
Sequence5  EDDEE
```

	1	2	3	4	5
1	x	0	2	5	3
2	0	x	2	5	3
3	2	2	x	3	5
4	5	5	3	x	2
5	3	3	5	2	x



Final tree with collapsed low confidence nodes



Bootstrap support for tree

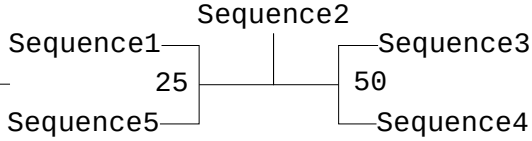


Figure VII

A tree represents only one possible path of evolution that might have given rise to the observed sequences. A tree is clear, simple and no conflicting information is displayed. For example, tree2 is preferred over tree1 in the likelihood analysis, but no reasons are given. Excluding column

5 from analysis, for example, makes tree1 the most likely solution. Similarly, if columns 2 and 4 are excluded, parsimony would return tree2 as the best. An estimate of how much confidence we can have in the various phylogenetic trees would be helpful.

Bootstrapping (Felsenstein 1985) (FigureVII) provides a way to estimate the amount of conflicting phylogenetic information present in a multiple sequence alignment, or how well the underlying data supports the individual nodes. Bootstrap analyses are performed by comparing trees generated from subsets of the original alignment with the original tree. Subsets are generated by statistical sampling of alignment columns with replacement. FigureVII shows the basic bootstrap procedure. Columns are selected at random from the original alignment and used to generate a replicate alignment of the same size. Due to random sampling a limited number of times, some columns may be represented multiple times, some not at all. A phylogeny is then inferred for this new alignment and compared to the original tree. Nodes present in both the original and replicate tree gain an increase in confidence. A bootstrap value of 50 means that the node in question was recovered in 50% of the replicates. For example, the node combining sequences 1 and 5 to the exclusion of sequences 2, 3 and 4 was recovered in only one of 4 replicates and the node combining sequences 3 and 4 to the exclusion of 1, 2 and 5 occurred twice. Between 100 and 1000 replicate trees are usually calculated to estimate bootstrap support for phylogenies.

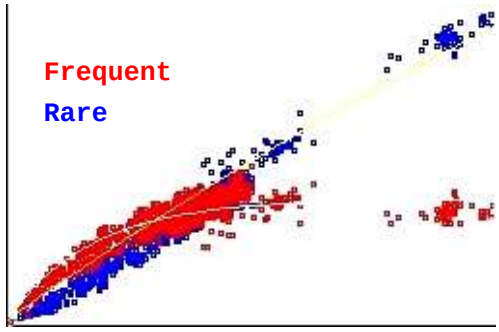
In general, groups with low bootstrap support are collapsed, causing a polytomy, the emergence of three or more branches from the same node. Polytomies indicate that further subdivision of that node is not supported by the data. In the “final tree” in FigureVII, the node separating sequences 1&5 from sequence 2 is collapsed due to low bootstrap support. This symbolizes that although we can predict a common ancestor for sequences 3&4, we are unable to say whether sequences 1&5, 1&2 or 2&5 are closest relatives.

Low bootstrap support for a node can have a multitude of reasons, but most frequently is due either to the bootstrap replicates being unable to resolve the tree or too much conflicting information in the alignment. In the first case, the alignment contains too few phylogenetically informative columns in relation to its length, increasing the probability that a statistical sampling of columns will miss a large number of them. This causes the replicate trees to greatly differ in topology and the bootstrap support for many nodes to drop. The second scenario is usually caused by either bad alignments or mutational saturation. Alignments are regarded as “bad” if residues that did not descend from the same position in an ancestral sequence are grouped in the same alignment column. Comparing nonhomologous features causes evolutionary scenarios to vary across replicates and thereby lowers bootstrap support for the tree. Mutational saturation refers to alignments in which multiple mutations at the same position have obscured the original phylogenetic signal. If, in the course of evolution, position X in sequence Y changed multiple times, for example from D>E, E>N, N>F, then what we observe is a change from D>F. Should speciation events have occurred during that time, common ancestry can no

longer be inferred from that residue. Preceding mutations cannot be recovered and some phylogenetic signal is lost. In a fully saturated alignment no phylogenetic information is retained. The observable sequence similarities are due to chance, random sampling of alignment columns causes large variation in the resulting phylogenies and bootstrap support is reduced.

Suboptimal alignment and mutational saturation.

Detecting mutational saturation of alignments is relatively straight forward and therefore less problematic to remedy. AsaturA (Van de Peer 2002) is a program that allows even highly saturated alignment to be used for phylogenetic inference. It uses information from substitution matrices to determine which amino acid exchanges are probable and therefore likely to occur frequently, and which are less probable and more rarely encountered. Plotting the number of "rare" and "frequent" substitutions over pairwise sequence dissimilarity produces a graph from which the alignment saturation can be estimated (FigureVIII). In this case, the graph shows a large amount of mutational saturation for "frequent" substitutions. For medium to distantly related sequences, center and right of the graph, no difference in the number of frequent substitutions is discernible. This points to the frequent substitutions being saturated and therefore unable to correctly reflect evolutionary history. For the rare substitutions, sequence divergence correlates in a linear fashion with the number of observable mutations. This points to negligible amounts of mutational saturation and makes these residues ideally suited for phylogenetic inference.



FigureVIII

Dealing with alignments containing nonhomologous sequences is more problematic. Global alignment programs do not verify whether or not the provided sequences are homologous, all they do is align sequences as best they can. Improving alignment quality therefore requires both checking the homology of aligned sequences and estimating the reliability of each alignment column. The latter can be achieved by comparing alternative alignment strategies and increasing the confidence for recurringly grouped residues. One way to derive estimates both for sequence homology and residue confidences is to use local alignments of all possible sequence pairs. Local alignments, such as generated by the sequence similarity search tools BLAST or PSI-BLAST provide both a statistical measure of sequence similarity and alternative alignments for all pairs of sequences. Residues aligned in the same fashion in both the global and pairwise alignments are increased in confidence, resides aligned differently are reduced in confidence. The amount by which the confidence changes is dependent on the P-value of the local alignment, i.e. the probability of the sequence

similarity being due to chance. The lower the P-value, the more the local alignment will influence alignment confidences. Using this approach it is possible to rate both sequences and individual residues as to how likely they are to be homologous compared to the rest of the alignment.

Once such estimates are available, sequences or alignment columns of low confidence can either be excluded from analysis or manually reexamined and improved.

Evolution, lateral-gene transfer and genome duplication:

Phylogenetic trees have been used throughout history to classify organisms according to shared ancestry. Some of the best examples for this may be found in the extensive and convoluted family trees of ancient European noble houses, which charted the appearance, birth, and extinction, death, of individuals over time. Similarly, phylogenetic trees have been used by biologists to visualize the presumed common ancestry as well as appearance and extinction of species.

Over the course of history, many scientists have attempted to order all living organisms according to some grand classification scheme. The undoubtedly most influential scheme was proposed by Carolus Linnaeus in his Systema Naturae in 1735. Even though our ideas about how the world came to be have changed dramatically, the basics of his proposal are still in widespread use, such as the binomial nomenclature and the hierarchical classification of groups within groups.

On July 1st 1858 manuscripts from both Charles Robert Darwin and Alfred Russel Wallace were presented to the Linnean Society of London and marked the first public airing of the hypothesis of natural selection. The idea of natural selection influencing the evolution of species gave classification schemes a new foundation. The ideas of descent with modification and natural selection provide the basis for the evolutionary theories currently in use. Phylogenetic trees attempt to combine hierarchical classification with evolutionary theory. A tree displays how, based on our understanding of evolution, larger and larger sequence groups may have emerged from a single ancestor, until the tree describes the world we observe today.

Studies attempting to reconstruct the tree of life, i.e. classifying all extant species, have historically been based on analyses of phenotypic traits shared between organisms. Zuckerkandl and Pauling showed in the early 1960's (Zuckerkandl 1962) that molecular sequences contained large amounts of information encoded in their characters, however, prior to the 1970's most phylogenetic analyses were still performed on anatomical similarities as these were the most easily available data and enabled the comparison of existing species with fossils. In the mid-1970's the advent of DNA sequencing methods made the world of molecular data readily accessible to phylogeneticists. DNA or protein sequences provided easy access to a large number of mostly independent traits, ideal as a basis for phylogenetic inference, and quickly spread in use. Surprisingly, however, trees inferred from nucleic or protein sequences did not always correspond to the accepted evolutionary history of an organism and were found to frequently contradict other gene trees.

In the 1990's Carl Woese proposed replacement of the bipartite tree of life containing prokaryota and eukaryota, with a tripartite representation in which bacteria, archaea and eukaryota emerge as major branchings, so as to reflect the profound differences molecular analyses had uncovered between these three groups (Woese 1990). These differences had not been apparent in phenotype based trees, due to the limited number of features comparable across microbes. Trees based on the small subunit ribosomal RNA (ssrRNA), a ubiquitously represented and highly conserved sequence, have since become the basis for our classification of distantly related organisms (Saccone 1995, Brochier 2002, Cavalier-Smith 2004). However, problems with the scenario of exclusive inheritance from parent to offspring (vertical inheritance) appeared in the mid-1990's as more molecular data became available and multiple sequence families were shown to produced trees in severe conflict with the ssrRNA phylogeny (Gupta, R.S. 1998, Doolittle 1998, Martin 1999).

These conflicting groups of sequences were at first thought to have been caused by extremely rare events of DNA exchanges of unrelated organisms and the resulting evolutionary implications to be of negligible importance compared to the effects of vertical inheritance. Over the past decade, however, the importance of lateral gene transfer (LGT), the transfer and incorporation of foreign genetic material in a genome, has been recognized and has revolutionized our theories regarding bacterial evolution.

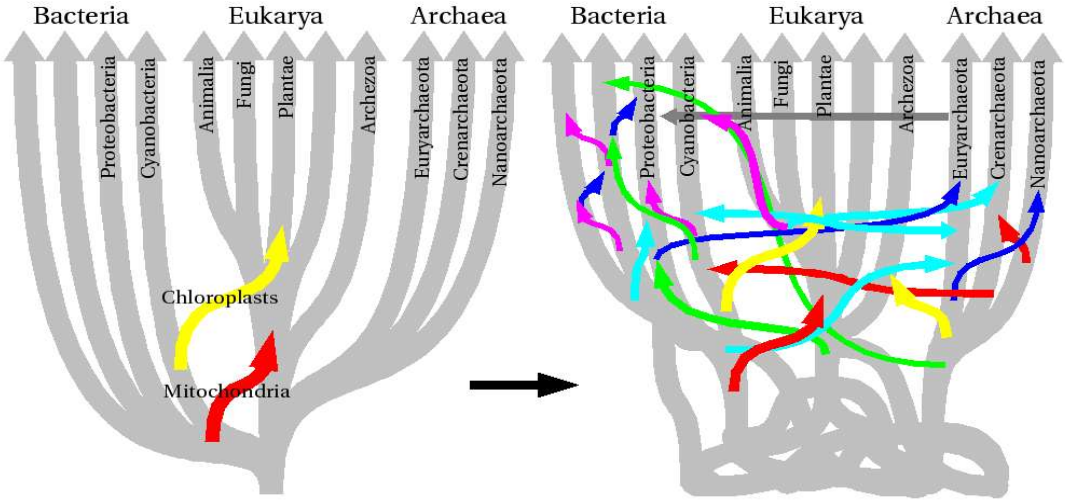


Figure IX Adapted from Doolittle 1999

In some cases LGT has been described as “the major, if not the sole, evolutionary source of true innovation: novel enzymatic pathways, novel membrane transporter capacities, novel energetics” (Woese 2000) and that it “maintains the universality of the genetic code... because the code is an evolutionary lingua franca required for an essential 'genetic commerce' among lineages” (Woese 2000). It has also been proposed that the extent with which LGT occurs has made it necessary to redefine the species concept as genes seem to be readily exchanged among many distinct lineages (de la Cruz 2000, Ochman 2000). Evidence for extensive amount of

lateral transfer have been found in multiple prokaryotic genomes, with LGT being offered as the most likely hypothesis for up to 24% of all open reading frames (ORF's) and even higher percentages in certain regions (Nelson 1999, Ruepp 2000, Cohen 2003).

This alternative method of gene propagation is, by now, firmly established as relevant to the evolution of bacterial genomes, but automated means of detecting LGT's in the course of genome annotation efforts have, so far, relied on pairwise sequence similarities even though they are known to be a suboptimal means of detecting sequence relatedness. It was only in 2001 that Sicheritz-Ponten and Andersson presented a tool for automated detection of LGT's in microbial genomes based on phylogenetic trees (pyphy) (Sicheritz-Ponten 2001).

LGT's, however, are not the only way of producing gene-trees in conflict with the species tree. One of the easiest ways of enabling novel evolutionary discoveries is to multiply the number of proteins selection can work on. This is achieved most frequently via duplication of either individual genes or larger genomic regions (Stephens 1951, Ohno 1967, Ohno 1970), the duplicate genes subsequently either evolving towards new functions or being lost. Differential gene duplication or loss can therefore cause trees to contain one gene for certain species and a number of homologs for others. Such trees can be regarded as contradicting the species-tree, as one species is represented by multiple sequences in the gene-tree.

Gene duplication, gene loss, lateral transfer and a number of other events influencing genome evolution can therefore be reconstructed by comparing gene trees to the phylogenetic history of a species and observing where, as well as the manner in which they differ. Phylome analyses attempt exactly this. The phylome represents the complete set of trees derived for the proteome of an organism. By comparing all individual trees in the phylome to the ssrRNA phylogeny, the gene accepted to best reflect the evolutionary history of species, it is possible to infer whether observed discrepancies occurred individually, for single genes, or if the event causing the differences encompassed a larger region, possibly the entire genome.

In 1976 Dingerkus and Howell proposed that a genome duplication gave rise to the ray-finned fish (actinopterygia), based on the large number of chromosomes found in species whose ancestors split off early in the evolution of this lineage (Dingerkus 1976). Multiplying the number of genes in the genome may have been a key event enabling further diversification and adaptation for this hugely successful lineage. A number of studies have since provided support for this hypothesis, such as the seven hox clusters found in Zebrafish (Amores 1998), the phylogenetic trees for 49 clades of orthologous proteins (Taylor 2003) and the analysis of the Tetraodon nigroviridis genome draft (Jaillon 2004). Just as phylogenies are the most accurate method of determining closest sequence relatives, they also provide the most accurate means of recovering the more distant evolutionary relationships of sequences. If many genes in a genome show identical duplication patterns and the events coincide with each other, it is plausible to predict a large duplication event occurring once rather than a number of smaller events occurring multiple times. Further evidence, such

as chromosome location and collinearity of regions can be used to corroborate or disprove such large duplication hypotheses. Taylor et al. performed a large scale analysis of the danio proteome using phylogenetic reconstructions to determine clades of orthologous genes in support of the duplication hypothesis. Unfortunately their method was based on manual generation and examination of large sets of trees which prohibits their approach from being routinely applied to other genomes. A tool able to automatically select sequence homologs, generate alignments, infer trees and automatically select those of relevance to the question at hand, would be a major facilitator for many comparative genomics analyses.

Protein classification:

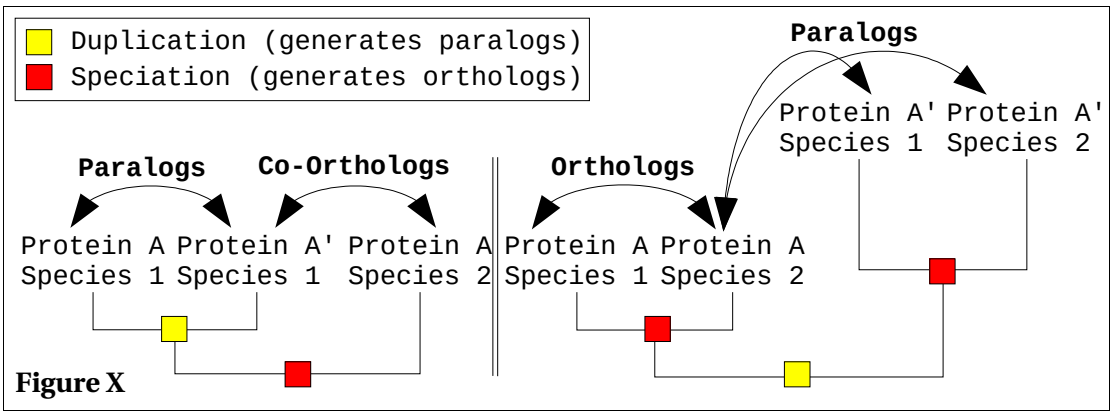


Figure X

While efforts in genome annotation and comparative genomics are laudable, the quest for knowledge about proteins can be extended in other directions as well. Large sequence families that arose via radiation of ancestral genes make it difficult to assign the correct ortholog for a new sequence and thereby greatly reduce the predictive capabilities of phylogenies. Sequences that arose via a speciation event are referred to as “orthologous” while sequences that arose through gene duplication are said to be “paralogous”. The description “co-ortholog” is given to genes that duplicated in one species but not another, as simple orthology would not account for the fact that either of the duplicates may since have evolved a novel or retained only part of the original function. Proteins under neutral selection, such as recent duplicates, one duplicate being able to counteract the effects of deleterious mutations in the other, accumulate mutations at a much higher rate and generally either evolve new functions or are removed from the genome. Standard duplication scenarios have three outcomes: A) one of the duplicates is lost; B) one of the duplicates evolves a novel function and both are retained; C) both duplicates specialize on performing different sub-functions of the original gene and thereby remain essential. Evidence for events combining radiation and subsequent selection can be found in a multitude of gene families. A nice example are the histidine kinases and response regulators, where different organisms multiplied different family members independently of one another to cover a wide variety of functions (Koretke 2000, Rabus 2004). Figure XI shows the histidine kinase and response regulator complement of Desulfotalea psychrophila, a cold loving bacterium living in arctic ocean sediment. The variously colored triangles

symbolize different families of histidine kinases. The size of each symbolizes the number of representatives present in the genome. The smaller triangles for the groups Sam, Syn, Anti-sigma, Mth, Lyt, Arf and PDK show that no sequence representatives for these families could be found. On the other hand, the Ntr family is excessively represented and, based on phylogentic reconstruction, can be subdivided into two subfamilies both for the histidine kinases as well as for the corresponding response regulators.

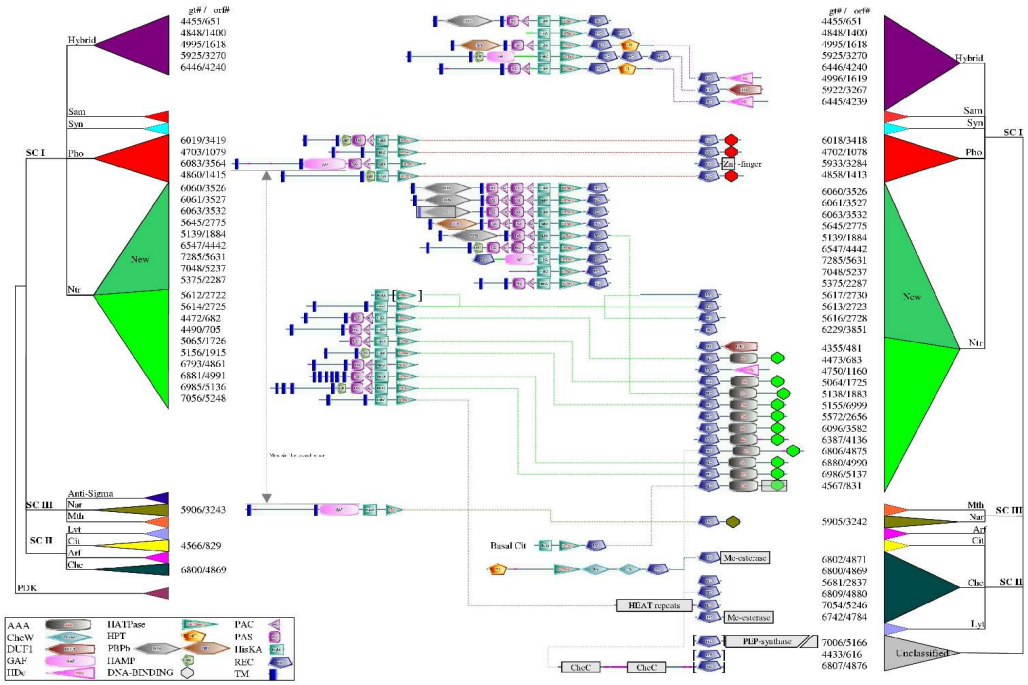


Figure XI (adapted from Rabus et al. 2004)

Simply because two organisms share an identical number of sequences from the same protein family, it cannot be deduced that each of the sequences has an ortholog in the other species. As phylogenies attempt to recover the evolutionary history of genes and are heavily dependent on multiple sequence alignments, problems at the alignment level, such as the difficulty of distinguishing between orthologs, co-orthologs and paralogs, especially pronounced if some of the paralogous genes have been lost, can greatly reduce the confidence estimates for phylogenetic reconstruction procedures.

A prime example for problematic classification is provided by the family of “ATPases Associated with diverse cellular Activity” (AAA-ATPases) first described in 1991 (Erdmann 1991). This highly diverse family is represented in all living organisms and its members are associated with a wide range of functions such as gene expression, vesicle mediated transport, membrane fusion, peroxisome and endosome biogenesis, proteolysis, microtubule severing and control of cell division. Although functionally diverse, family members commonly form hexameric rings and unfold proteins in an energy dependent manner. The multidomain nature of these proteins, consisting of

a N-terminal domain followed by one or two ATPase domains, complicates the analyses. On the one hand, the homology of N-domains of many sequences is unclear, on the other hand, AAA proteins with two ATPase domains may contain one inactive domain, for example domain 1 in peroxisomal AAA-proteins and domain 2 in NSF (N-ethylmaleimide-sensitive factor) or have retained two active versions such as p97/CDC48. Inactive domains are under less selective pressure, accumulate mutations more rapidly, make the task of finding and aligning homologous active domains more difficult, introduce noise and therefore should be removed prior to phylogenetic inference. Mutational saturation of the dataset further complicates the picture. The presence of AAA-ATPases in all living organisms suggests that these proteins were already present at the time of the last universal common ancestor, the organism that gave rise to all living organisms. As a AAA-ATPase domain is only approximately 240 residues long, it can be expected that some of the limited number of mutatable positions, those for which a residue change will not inactivate the protein, will have changed multiple times since the divergence of bacteria, archaea and eukaryota. Subsequent mutations at the same position mask phylogenetic information, introduce noise and complicate the process of tree inference.

The ancient evolutionary origin, numerous sequence representatives, mutational saturation, differentially selected or inactivated domains and highly diverse functions make this family both interesting and challenging to classify. Even though this sequence family has been studied for over a decade, a multitude of representatives are known and a number of major subfamilies are readily definable, deeper insights into the evolutionary history of AAA-ATPases are not readily apparent. Various attempts at classification have differed in the set of sequences used, the approach and in the treatment of two-domain AAA-ATPases. (Beyer 1997, Swaffield 1997, Wolf 1998, Froehlich 2001, Lupas 2002). A clear delineation of the family members and robust estimates for the basal branching pattern would provide a sound basis for future research into history, major evolutionary events and mutational constraints of the AAA-protein family.

Aim:

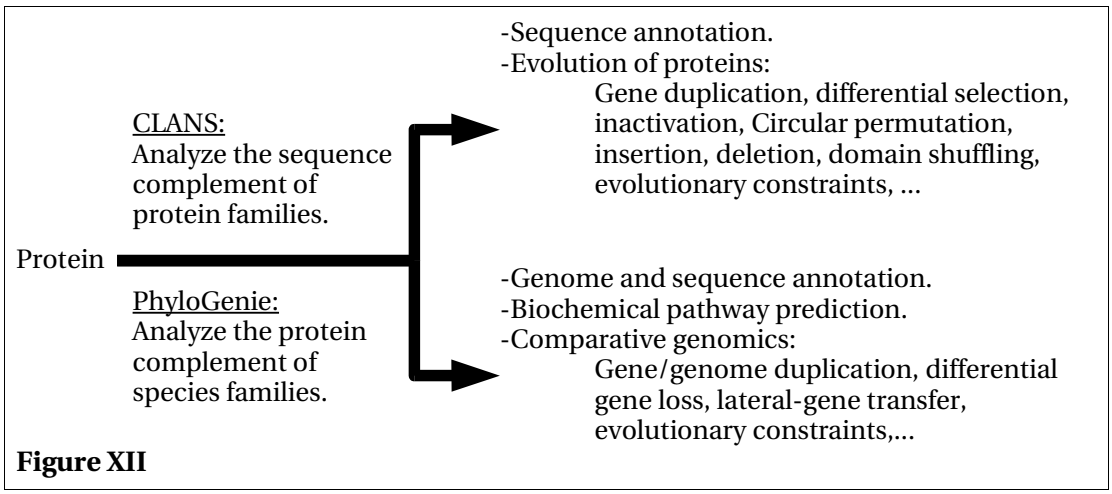


Figure XII

The amount of data being generated by the various genome projects is enormous and the direct cause of a number of problems: A) Databases are flooded with sequences containing “hypothetical” or “unknown function” as sole annotation; B) sequence data is often of low quality and subject to frequent changes; C) database size has grown exponentially, causing similarity search programs to return large numbers of potential sequence homologs that need to be examined.

Tools able to analyze large amounts of data are mostly inadequate at classification and phylogenetic inference methods have severe difficulties when faced with large datasets. Improving the resolution of a phylogeny can generally be done by either extending the alignment to include more phylogenetically relevant residues, or by adding more sequence intermediates. Every additional sequence increases phylogenetic signal as well as noise. Depending on sequence diversity and alignment length, every alignment reaches a point where increasing the number of sequences will introduce more noise than phylogenetic signal. Adding more sequences beyond that point only exacerbates the problem. Therefore, all multiple sequence alignments have an upper limit to the number of nodes they can resolve via phylogenetic inference. In addition, phylogenetic trees are a lot more difficult to analyze in an automated manner than BLAST similarities.

Comparative genomics frequently rely on BLAST results as their main data source due to difficulties in automating phylogenetic inference and analysis. For protein family analyses, the major problem lies in the large number of potential sequence homologs. Although it is possible to base a phylogeny on a representative sample of the family, determining that sample is difficult. Classification of the family is needed for representative sampling and a representative sample is required for classification.

On the one hand we attempt to develop methods that analyze whole proteomes in an automated manner, thereby extending the applicability and ease of use of comparative genomics to everyday biological problems. On the other hand we look for classification schemes that are capable of handling the large number of sequences returned by similarity searches for some of the larger protein families.

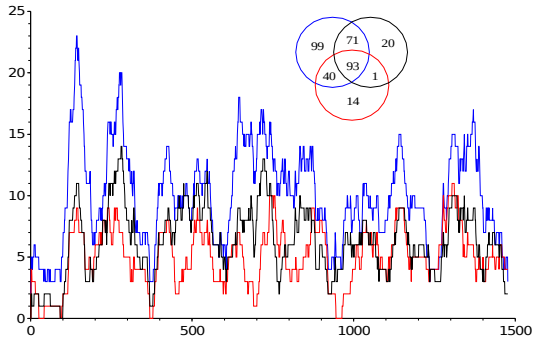
Although methods development is a central part of this work, we want to place at least as much emphasis on application to biological data. For this reason we decide to repeat three analyses in which we believe previous methodological insufficiencies may have hindered robust and in-depth assessments: I) Detection of all lateral-gene transfers between Thermoplasmata (*Thermoplasma acidophilum*, *Thermoplasma volcanium*, *Ferroplasma acidarmanus* and *Picrophilus torridus*) and Sulfolobus (*Sulfolobus solfataricus* and *Sulfolobus tokodaii*), II) Detection of all zebrafish genes supporting the actinopterygian specific genome duplication hypothesis and III) Classification of all AAA-proteins.

Results and Discussion:

The first aim was to analyze lateral-gene transfer events in bacterial genomes. Although a reasonable amount of work had already been performed in this field, a rapid overview of the available software showed a pressing need to extend upon existing tools to enable more efficient analysis. Alternative methods, such as best BLAST hits or "pyphy" (Sicheritz-ponten and Andersson 2002), appeared as suboptimal solutions to the problem. Best BLAST hits had previously been shown to produce extensive amounts of false-positive predictions and the pyphy sequence selection and alignment procedures seemed inflexible, excessively stringent and easy to improve upon. Once the basic problems concerning sequence selection, alignment and tree inference were resolved, a means of efficiently searching large numbers of trees for those with interesting topologies was developed, as manual examination of hundreds of trees seemed neither the most rapid nor the most efficient way of gathering data (Frickey & Lupas 2004). To test the validity of our approach and the applicability of our tools, we examined two datasets: I) the Thermoplasma acidophilum genome and II) the incomplete Danio rerio genome, and compared our results to previously published analyses (Ruepp et al. 2000, Taylor et al. 2003).

Project I: PhyloGenie (Frickey T and Lupas AN. 2004, Nucleic Acids Res.)

The *Thermoplasma acidophilum* genome was examined for lateral-gene transfer events between the two distantly related archaeal lineages "Thermoplasmata" and "Sulfolobus". The LGT analysis recovered a large number of clades of orthologous proteins in which Thermoplasmata and Sulfolobus representatives were closest relatives. The distribution of these genes over the *Thermoplasma acidophilum* genome supported the claim of a few transfers of large DNA regions between these two lineages, although with a lower overall amount of genes involved than predicted by Ruepp et al. (2000). The transfer of large genome fragments from one organism to another, coupled with the procaryotic tendency to group functionally linked genes in a genomic region, explains how complex metabolic pathways combining numerous interdependent genes are able to appear suddenly in prokaryotes previously lacking the feature.



Analysis of the Danio rerio proteome for all proteins supporting the hypothesis of a fish-specific genome duplication, greatly improved upon previous analyses. Taylor et al. 2003 used a BLAST based approach followed by manual alignment, tree inference and analysis to identify 49 clades of orthologous groups showing a 2:1 ratio of zebrafish to tetrapod genes. Our reanalysis of the zebrafish proteome was able to increase the number of clades in support of this hypothesis to 120 in a fraction of the time required by the Taylor et al. approach. Interestingly, most of these clades were of

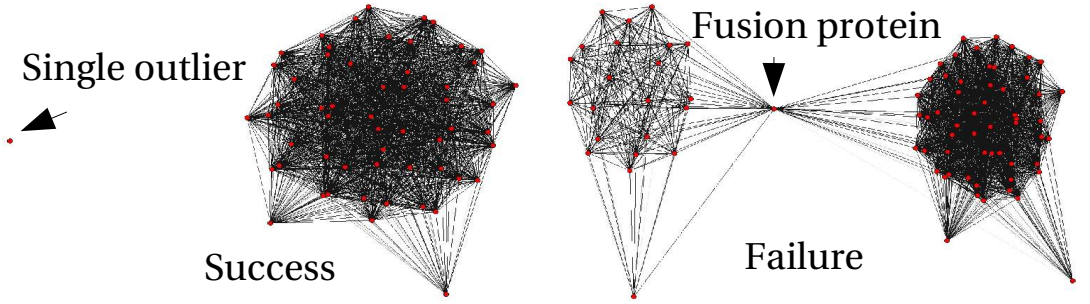
morphogenic or regulatory nature. This effect may have been caused by either A) massive loss of non-morphogenic gene duplicates, or B) as neither the zebrafish nor the red junglefowl genomes were completed, the bias may simply indicate a historical preference of molecular geneticists to study morphogenic and regulatory genes.

Direct comparison with the most frequently used LGT detection methods, BLAST and pyphy, showed that our tool was superior to both. BLAST based results contained far too many false positives, up to 40% of all predictions, while pyphy proved to be both less specific and less sensitive than our approach. In addition, neither BLAST nor pyphy are able to analyze questions about the evolution of proteins encompassing more than the closest sequence relatives, thereby severely limiting the range of testable hypotheses. PhyloGenie allows for queries containing complex sequence relationships and promises to greatly extend the use of phylome analyses in future comparative genomics studies.

Project II: Alignment validation

As the performance discrepancy between pyphy and PhyloGenie was predominantly based on differences in quality of the alignments generated by the programs, we decided to further refine our alignment procedure. A lot of work had already been invested in the steps regarding sequence selection and alignment, causing us to focus on post-processing the available alignments.

Estimating alignment quality is a two-step process: first, the homology of all aligned sequences has to be ascertained and second, confidence values for individual alignment columns have to be calculated. We chose to base our confidence calculations on BLAST or PSI-BLAST pairwise local alignments. Residues aligned in the same fashion both by global and pairwise, local alignment programs were increased in confidence; residues aligned differently were reduced in confidence. The BLAST hit P-value determined the amount confidences changed by. Testing this approach on a few alignments retrieved from the SYSTERS database (Krause et al. 2000) and modified for our purposes, revealed a severe problem. While capable of detecting single false positives in an alignment, the method failed whenever presented with multiple groups of nonhomologous sequences. FigureXIII shows both examples. The left panel shows a graph representation of sequence similarities for a case where the validation procedure worked. The single outlier sequence was detected as nonhomologous and assigned a correspondingly low alignment confidence (data not shown). The right panel shows the more complicated case of two nonhomologous sequence families present in one alignment. The single sequence with connections to both groups was a man-made fusion protein and the reason why both groups appeared in one alignment. In this case each group self-validated independently of one another and the complete alignment was assigned high-confidence values although clearly containing nonhomologous features. Fortunately these cases could be detected by extending the alignment validation procedure with a force directed layout of all pairwise sequence similarities. The clusters apparent in the layout allowed us to



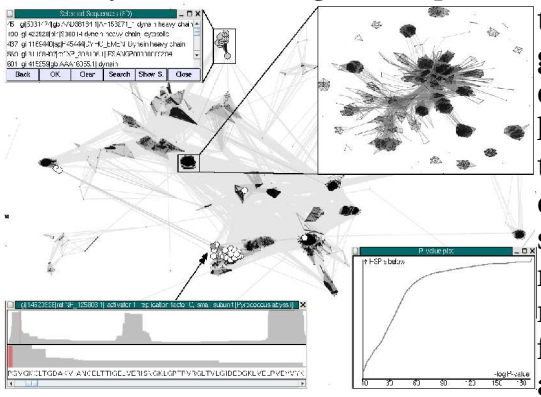
FigureXIII

automatically recognize whenever multiple families were present in an alignment and correspondingly adapt the validation strategy.

The force directed placement method was subsequently greatly extended and used in a number of analyses as it provided a means of rapidly gaining an overview of groups of unaligned sequences. Contrary to phylogenetic trees, the inclusion of nonhomologous sequences seemed to have little effect on the clustering and resolution improved with increased number of sequences.

Project III: Clans (Frickey T and Lupas AN. 2004, Bioinformatics)

The program CLANS represents a refined version of the force directed clustering approach, optimized for sequence family detection and analysis. Its ability to use unaligned sequences and work with datasets too large for traditional phylogenetic inference greatly extend upon existing capabilities for rapid classification of large protein families. In addition, as the approach is insensitive to the detrimental effects nonhomologous sequences have on phylogenetic reconstructions, it provides a useful means of separating true homologs from chance hits prior to sequence alignment and tree inference.



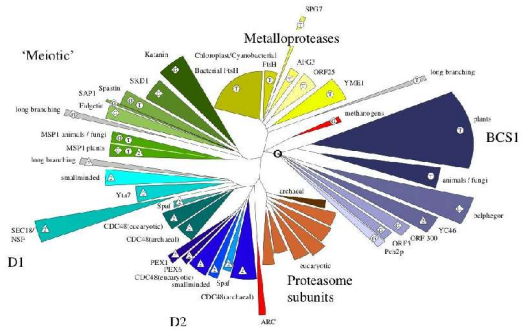
The analytical capabilities of this tool include, but are not limited to, automated detection of sequence families, estimating their robustness, determining the regions containing family sequence signatures and the ability to focus on and examine selected sequence subsets in greater detail.

This tool played a pivotal role in our ability to clearly delineate and classify the AAA-ATPase protein family. Although previous phylogenetic analyses always included a core group of AAA-proteins, a number of sequence families were included in some but not others and no analysis gave objective reasons for why their specific set of sequences was supposedly representative for AAA-proteins. Additionally, the inconsequent treatment of sequences, some being excluded from analysis as too divergent although containing all canonical residues and some being included

although missing the required catalytic residues, prompted us to search for a more objective approach.

Project IV: Phylogenetic analysis of AAA-ATPases (Frickey T and Lupas AN. 2004 , J. Struct. Biol.)

An extensive search of the non-redundant NCBI protein sequence database for putative AAA-ATPases returned 5101 proteins. A CLANS based analysis showed that this set contained representatives for all major families present in the AAA+-superfamily and that all sequences known to belong to AAA-proteins were present in a well-defined group clearly distinguishable from the rest. Having representatively sampled the AAA+-superfamily in a search centered solely on AAA-proteins, increased our confidence in having exhaustively sampled the AAA-family and thereby identified all sequences present in the database. Phylogenetic analysis of the ATPase-domains for the AAA-group recovered all of the previously defined AAA clades and, additionally, showed clades omitted from some analyses to be basal members of the AAA-family. Saturation correction was used to improve resolution of the tree and better define the basal branching order.



We showed that detection and classification of all AAA-family members in the NCBI non-redundant protein database was possible and extended upon previously described evolutionary scenarios by defining the family, focusing on the basal branching pattern and providing a tentative root for the tree. Correcting for mutational saturation of the sequence alignment resolved some of the problems observed in other phylogenetic reconstructions, such as the grouping of domain 1 of NSF with domain 2 of other two-domain AAA-ATPases and thereby increased confidence in the general correctness of the tree. Assuming the tree to accurately reflect evolutionary history, a number of surprising conclusions are apparent. Two domain ATPases seem to have arisen on at least three independent occasions, once in the ancestor of “traditional” two-domain ATPases (CDC48, NSF, PEX, etc.), and once each in the clades MSP1(plants) and YC46. Although unexpected, the polyphyly of two-domain AAA-ATPases is acceptable, as it is due solely to new sequences containing degenerate second domains having been detected in the two latter groups. A further surprise was that the second region of homology, a defining characteristic of all AAA-ATPases, seems to have repeatedly lost the two residues immediately preceding the “arginine finger”. Additional minor clades, mostly procaryotic and for which little is known, were shown to be basal members of the AAA-family. Analysis of the N-terminal domains provided some unexpected results, such as a deep split between plant and animal YME1 proteins, or the apparent homology of two distantly related groups such as CDC48 and archaeal methanogens or ARC and proteasomal ATPases. The results obtained were biologically plausible

and unexpected predictions arising from the cluster analysis, such as the apparent homology of N-domains of distantly related AAA-proteins, were subsequently corroborated by other methods. Taking this tree as a basis, some of the ideas about the evolution of this protein family may have to be revised.

As a side effect, force directed placement was shown to be an effective means of preliminary classification for large datasets. Contrary to phylogenetic inference methods, there is no inherent upper limit to the number of sequences this method can work with and, as larger numbers of pairwise comparisons better average out false positive random similarities, classification improves with increased number of sequences. In addition, the method uses unaligned data, and is capable of identifying nonhomologous sequences. The graph representation of all pairwise similarities is less abstract than a phylogenetic tree and thereby better approximates the actual data. Even though this method provides a better representation of sequence similarities and is able to rapidly classify large datasets, it is unlikely to replace phylogenetic reconstructions. Phylogenetic trees group together sequences along one axis, evolutionary time, and thereby generate a hypothesis about how individual sequences are related to one another. Force directed placement does not attempt to infer evolutionary relationships. The only aim is to rapidly find a reasonable representation of the underlying data, providing a basis from which the most relevant sequences can be selected.

The successful use of the clustering based approach for the AAA-dataset led us to apply it in a number of other analyses. As a direct extension of the above work, we classified the TAA43 protein of Thermoplasma acidophilum as belonging to a group of archaeal AAA-ATPases most closely related to the “meiotic” AAA-proteins found in eukaryotes (Santos et al. 2004). An analysis of Wipi-1-alpha also benefited from cluster analysis, as it enabled us to determine a small group of closely related sequences from within a large subset of WD40-proteins. Based on this smaller group, we were able to predict two duplication events in evolutionary history of the family, putative catalytic residues and a probable function for the protein (Proikas-Cezanne et al. 2004). CLANS is also being applied to ongoing analyses, such as examination of the extended family of AbrB transcription regulators (Djuranovic S., Coles M., MPI Tuebingen, unpublished), TIM-barrel proteins (Sergeev Y., MPI Tuebingen, unpublished), visualization of the similarities of very distantly related protein families (Soeding J., MPI Tuebingen, unpublished) as well as analysis of the families of two-component signal transduction proteins (TCST) (Figueroa F., MPI Tuebingen, unpublished). Comparison of TCST family assignments, based on either manual analysis of multiple sequence alignments, phylogenetic trees or CLANS, shows all three methods recovering the same families. These results provide a further example for our clustering approach being able to correctly classify a protein family.

Contribution:

The numerous discussions I have had with Andrei Lupas throughout the course of this PhD. make it difficult to disentangle the individual ideas each of us contributed. Therefore, unless stated otherwise, joint discussions between Andrei Lupas and myself are to be regarded as having generated the ideas presented herein.

PhyloGenie:

Development of PhyloGenie was a two step process. During my studies at the Constance University I was involved in the Taylor et al. (2003) analysis regarding the ray-finned fish specific genome duplication. The large number of simple and repetitive steps the analysis required, first prompted the idea for such a program.

After beginning my PhD. at the Max-Planck Institute for Developmental Biology in Tuebingen, I was faced with the task of developing a tool to detect all lateral gene transfers in a given genome. This involved automating the selection of sequence homologs, generating alignments and inferring and analyzing phylogenetic trees. Andrei Lupas and Kristin Koretke had already done some work on how to best convert BLAST results to multiple alignments, therefore the alignment scheme should be attributed to them, but no framework integrating homology search, alignment, tree inference and analysis was present.

All programs were conceived and written by myself; all analyses, data and figures presented are my own work.

CLANS and AAA-ATPase analysis:

Problems in automated selection of sequence homologs and alignment prompted the development of CLANS. The first use was the detection of nonhomologous sequences in multiple alignments. It rapidly became clear that the method could be used to classify sequences according to their respective families and the idea of analyzing datasets too large to handle for traditional classification schemes was born.

Andrei Lupas provided me with an introduction into the AAA-family, numerous insights regarding the relationship of individual AAA subgroups and countless helpful discussions. The program CLANS was conceived and written by myself; all analyses, data and figures are my own work.

Appendix:

BLOSUM 62 substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Published online September 30, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 17 ***5231–5238***
doi:10.1093/nar/gkh867

PhyloGenie: automated phylome generation and analysis

Tancred Frickey and Andrei N. Lupas*

Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Spemannstr. 35, D-72076 Tuebingen, Germany

Received May 14, 2004; Revised July 29, 2004; Accepted September 15, 2004

ABSTRACT

Phylogenetic reconstruction is the method of choice to determine the homologous relationships between sequences. Difficulties in producing high-quality alignments, which are the basis of good trees, and in automating the analysis of trees have unfortunately limited the use of phylogenetic reconstruction methods to individual genes or gene families. Due to the large number of sequences involved, phylogenetic analyses of proteomes preclude manual steps and therefore require a high degree of automation in sequence selection, alignment, phylogenetic inference and analysis of the resulting set of trees. We present a set of programs that automates the steps from seed sequence to phylogeny and a utility to extract all phylogenies that match specific topological constraints from a database of trees. Two example applications that show the type of questions that can be answered by phylome analysis are provided. The generation and analysis of the *Thermoplasma acidophilum* phylome with regard to lateral gene transfer between Thermoplasmata and Sulfolobus, showed best BLAST hits to be far less reliable indicators of lateral transfer than the corresponding protein phylogenies.The generation and analysis of the *Danio rerio* phylome provided more than twice as many proteins as described previously, supporting the hypothesis of an additional round of genome duplication in the actinopterygian lineage.

INTRODUCTION

The amount of sequences being generated by genome projects far exceeds our ability to manually assign any meaningful annotation to them. To analyze the flood of 'unknown' or 'hypothetical' sequences in a reasonable time frame, automated methods are essential. These often rely on the assumption that sequences have the same function as their closest relative. The use of best BLAST hits to find these close relatives may often be a viable option (1). However, Koski and Golding showed that best BLAST hits do not necessarily represent the closest sequence relatives (2), thereby casting doubt on the reliability of this approach. The human genome consortium (3), for example, predicted 113 lateral gene transfers (LGTs) from bacteria to vertebrates based on BLAST results. Subsequent phylogenetic analysis of the genes in question, however, was unable to find support for any of these predictions (4–6).

The use of the trees to find the closest relatives, by inferring a phylogeny for each sequence, is a more robust but computationally demanding approach. It is difficult to automate reliably, as it involves two steps—selection of homologous sequences and multiple alignment—whose automated forms are error-prone. A program that automates the steps of similarity search, alignment and phylogenetic inference, Pyphy (7), uses a reduced sequence database with higher-quality annotation [Swissprot + TREMBL,(8)], fixed criteria of similarity to define homology (80% coverage and 50% identity, or identical annotation) and alignment of full-length sequences [ClustalW (9)]. Pyphy was specifically designed to detect and visualize LGT in prokaryotic genomes, and its restrictive settings were chosen to optimize its performance on this problem.

We have developed a suite of programs, PhyloGenie, which also automates the steps from seed sequence to phylogenetic inference, but can be used to examine a much broader range of phylogenetic hypotheses. PhyloGenie can be used with any standard FASTA format database, is based on local alignments, offers full flexibility in setting the criteria for homology and filters phylomes for all trees matching specific, user-defined topological constraints. To illustrate its operation and scope, we apply PhyloGenie to two phylogenetic problems that have been studied previously by non-automated methods and compare its performance with Pyphy. The two problems are the apparent large-scale LGT between *T.acidophilum* and *S.solfataricus* (10), two phylogenetically distant Archaea that inhabit the same environment, and the presumed additional genome duplication in the actinopterygian lineage since its divergence from tetrapods (11).

METHODS

Genomes and databases

NCBI taxonomy files and the non-redundant (nr) sequence database were obtained from the NCBI website (www.ncbi.nlm.nih.gov). The complete genome of *T.acidophilum* and all

*To whom correspondence should be addressed. Tel: +49 7071 601 340; Fax: +49 7071 601 349; Email: andrei.lupas@tuebingen.mpg.de

sequences for *Danio rerio* in the nr database of October 2003 were obtained from the same source.

Sequence similarity detection and alignment

Sequences were compared with the nr sequence database using BLASTP v2.26 and multiple sequence alignments were derived using the Java program Blammer. Blammer consists of five post-processing steps for BLAST result files that convert sets of high-scoring segment pairs (HSPs) to multiple alignments; this routine relieves the gapping problems that arise during the conversion of pairwise alignments to multiple alignments (Figures 1 and 2). All parameters (X to P) specified below can be customized and were chosen so as to maximize the number of BLASTP hits while providing reasonable support for sequence homology.

First, full-length sequences for HSPs up to expectation values (*E*-values) of X (X = 10) are extracted, which enables the sequence database to be searched with a profile hidden Markov model (HMM) (12) in a later step. The HSPs of the query sequence with a coverage greater than Y (Y = 60%) and *E*-values better than Z ($Z = 10^{-5}$) are extracted and the most dissimilar K (K = 150) of these are converted to a multiple alignment. The coverage and cutoff *E*-value are used to determine sequence homology and the most dissimilar HSPs are used to ensure that the HMM generated from the resulting alignment in a later step is representative of all of the relevant BLAST hits instead of only a large group of mostly identical sequences. Alignment regions with more than L (L = 100) consecutive ungapped columns are taken as alignment anchor points and all residues between such anchors are realigned using ClustalW, thus resolving inconsistent gapping problems.

A HMM, derived from the resulting alignment, is used to search the database of full-length sequences generated in the first step. This removes false positive BLAST hits and better defines the beginning and end of alignable sequence regions due to the higher sensitivity of HMMs. The alignment from which phylogenies are inferred consists of the HMM-HSPs with *E*-values better than M ($M = 10^{-6}$).

Sequences of the same organism with more than N (N = 99%) sequence identity are thought to be redundant database entries and only one copy is retained. In cases where the HMM search returns more than P sequences (P = 150), only the best P matches are converted to a multiple alignment so as to the keep ensuing phylogenetic calculations and analyses in a reasonable time frame.

Pyphy

The program Pyphy was obtained from T. Sicheritz-Ponten (Technical University of Denmark, Lyngby) and installed under Gentoo Linux. To make the output of Pyphy comparable

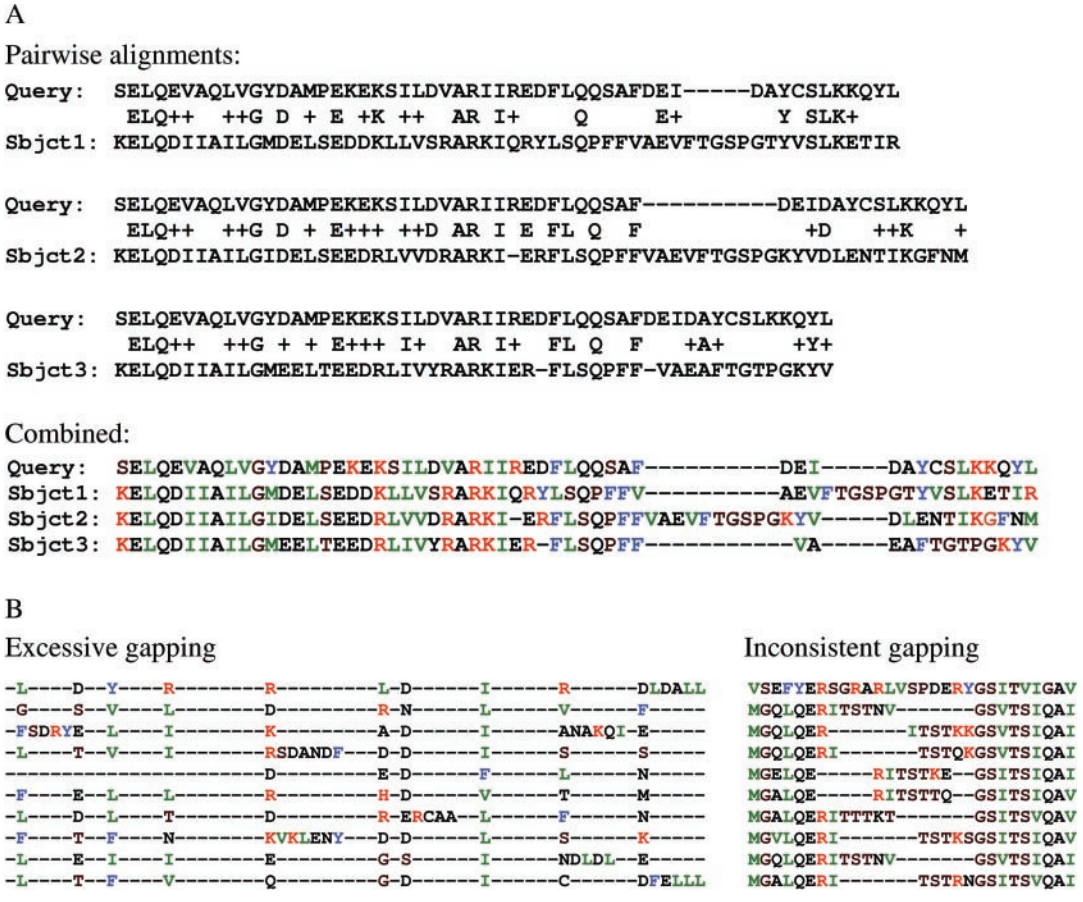


Figure 1. Alignment excerpts showing the most commonly encountered problems when converting BLAST or PSIBLAST HSPs to multiple alignments. (**A**) Three BLAST HSPs combined to a multiple sequence alignment and the resulting gapping problems. (**B**) Extreme examples of excessive and inconsistent gapping.

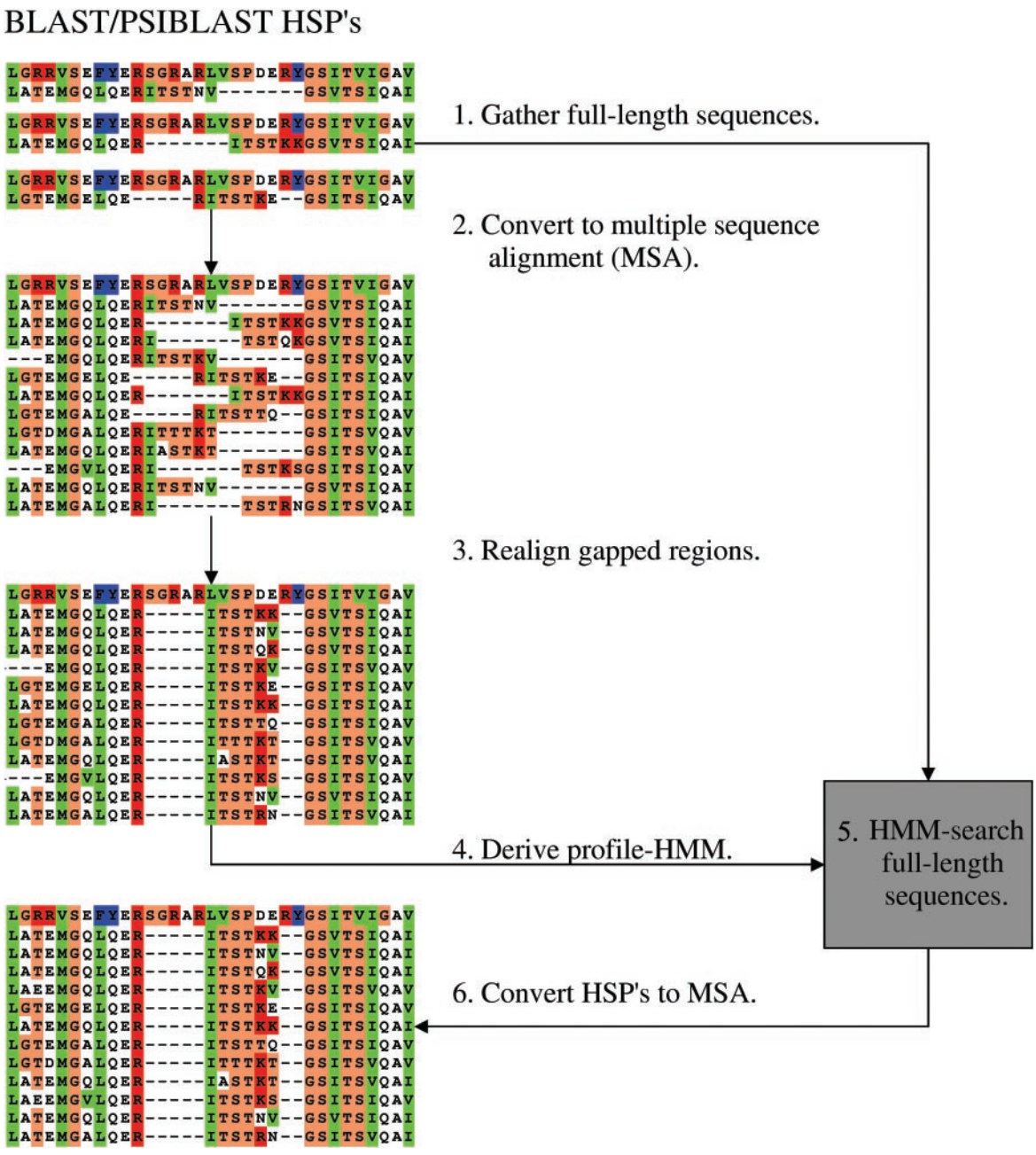


Figure 2. Layout showing the BLAST/PSIBLAST post-processing steps used to reduce excessive and inconsistent gapping. (1) All full-length sequences are gathered for HSPs and form the database used for HMM-searching in 5. (2) All HSPs matching E-value, score and coverage cutoff criteria are converted to a multiple sequence alignment. (3) The alignment sequences are filtered by maximum sequence identity to remove duplicate entries and gapped regions are realigned to resolve gapping problems. (4) A profile-HMM is derived from the multiple sequence alignment. (5) Sequences from step 1 are searched with the HMM generated in step 4 so as to better define the start and end of alignable regions and thereby improve alignment. (6) HMM-HSPs are converted to a multiple sequence alignment.

with PhyloGenie (specifically, to avoid distance versus parsimony issues), tree inference was handled in the same way for both programs by using the PhyloGenie routines.

Phylogenetic inference

Phylogenies were inferred using neighbor-joining (NJ) (13) in combination with the Poisson distance correction scheme and bootstrapped with 100 replicates.

External programs

For full functionality, it is necessary for the NCBI taxonomy files 'names.dmp' and 'nodes.dmp' (necessary for tree analysis) as well as BLAST (www.ncbi.nlm.nih.gov), HMMER (http://hmmer.wustl.edu) and ClustalW (www.ebi.ac.uk/clustalw) to be installed. To further customize the utility, it is possible to replace the alignment and tree construction routines. Any program or script that accepts FASTA format

sequences as input and generates clustal format alignments can replace ClustalW as an alignment tool. Similarly, any tree construction program that accepts aligned FASTA format sequences and generates Newick format trees can replace the provided NJ tool.

Tree analysis

The *T.acidophilum* phylome was searched for trees showing LGT between Thermoplasmata and Sulfolobus using the query '(Thermoplasmata & Sulfolobus & !(*cellular organisms))'. Trees corresponding to this search string included those with at least one node containing Thermoplasmata and Sulfolobus sequence representatives but no other cellular organisms.

For the zebrafish set of trees, the query '((Danio rerio {=2} & Homo sapiens {=1} & Mus musculus {=1} & Gallus gallus {=1} & Euteleostomi) & !(*Eukaryota))' returned phylogenies containing nodes in which two genes were present in *Danio rerio* and exactly one in *Homo sapiens*, *Mus musculus* and *Gallus gallus*. In addition, sequences belonging to non-euteleostomi eukaryotes were not permitted in that node.

Prior to analysis, sequences belonging to the NCBI taxonomic groups 'Viruses', 'Viroids', 'other sequences' and 'unclassified' were excluded and all nodes supported by bootstrap values below 50 were collapsed.

The analysis of unrooted trees is far more complex than that of rooted trees due to missing directionality (Figure 3a). However, automated rooting of trees is non-trivial. We have implemented the following rooting scheme that ensures correct directionality for at least the branch containing the seed sequence, i.e. the one the tree was calculated for, and frequently the complete tree. A tree is rooted by assigning a taxonomic 'level' to each node and rooting at the node with the lowest level (i.e. closest to 'root of life' or 'root') (Figure 3d). To assign a node's taxonomic level, the tree is first rooted with the seed sequence (Figure 3a: MAN) and the lowest common taxonomic denominator for all descendant species is calculated for each node (Figure 3b). Next, the tree is rooted at the leaf-node, the least related and having the highest number of nodes separating it from the seed sequence (Figure 3b: *E.coli* K12). All nodes are then reassigned a taxonomic level. If a node's new taxonomic level differs from the previous assignment, the level closest to 'species' is retained (Figure 3c). The second rooting and round of taxonomic assignments is done to remove directionality from the taxonomic assignments and ensure that they are

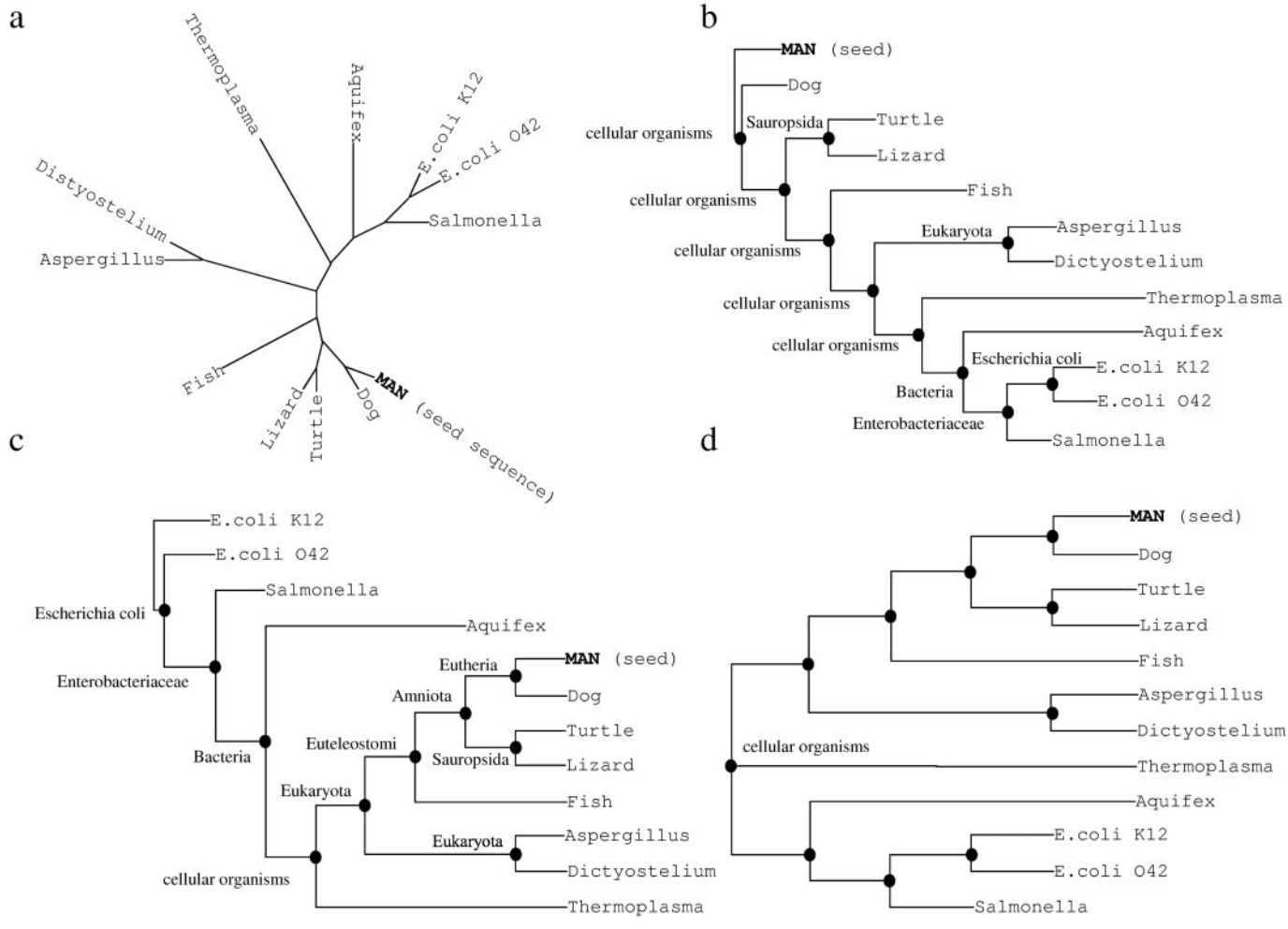


Figure 3. Tree rooting scheme. (**a**) Unrooted tree. (**b**) Tree rooted at the seed sequence (Man) with taxonomic ''level'' assignments for each node. (**c**) Tree rooted at the tipnode least related and most distant from the seed sequence (counting nodes) after the second round of taxonomic assignment. (**d**) Final tree, rooted at the most basal node the most distant from the seed sequence.

independent of the way the tree was rooted. The node closest to 'root of life' (last common ancestor for all proteins in this tree) is used to root the tree (Figure 3d). If multiple nodes of the same 'lowest' taxonomic level exist, the tree is rooted at the node most distant from the seed sequence.

Computing resources

The *T.acidophilum* analysis was performed on an AMD64 2400 1CPU workstation running Linux. Analysis of the *Danio rerio* proteome was done on a SUN V880 under Solaris9. All Pyphy analyses were performed on AMD64 2400 workstation running Linux. Generation of the *T.acidophilum* phylome required 78 h. The BLAST searches for each protein took 14 h, the conversion of BLAST to multiple alignments took an additional 60 h, and 4 h were needed to infer phylogenetic trees and bootstrap each with 100 replicates. The analysis of the resulting phylome took 36 s.

Availability

The software can be downloaded from http://protevo.eb.tuebingen.mpg.de/download.

RESULTS AND DISCUSSION

The PhyloGenie program

Analysis of phylomes, defined as the complete set of phylogenetic trees derived from the proteomes of organisms (7), requires four key steps: selection of homologs, multiple alignment, tree inference and filtering for specific tree topologies.

In Pyphy, the selection of homologs is guided to a large extent by sequence annotation. This requires high-quality sequence databases that provide standardized annotation, such as Swissprot and TREMBL, which prevent the use of most public databases. Since both Swissprot and TREMBL lag substantially behind the nr sequence database, both in number of sequences and completeness, we have implemented a sequence selection routine in PhyloGenie that is completely driven by local pairwise similarity. First, we extracted sequences with domain-sized regions of statistically significant sequence similarity, using the search tools BLAST or PSIBLAST, which are fast, reliable and sensitive. We then refined this set during the alignment process, using HMMs (see Methods; Figure 2).

Good phylogenies require good alignments. The errors incurred in the alignment process cannot be corrected by the subsequent steps of analysis. Non-homologous sequences or domains in an alignment, misaligned residues or the unfortunate selection of sequence representatives can lead to errors and possibly invalidate the inferred tree. Generating high-quality sequence alignments can therefore be seen as the most critical step on the path from seed sequence to phylogeny. When producing alignments, it is necessary to decide between aligning full-length sequences and aligning only the conserved regions for which sequence similarity, presumably due to shared descent, is unambiguously determinable. Pyphy uses the global alignment program ClustalW to align full-length sequences, thus requiring that all sequences in the alignment match over most of their length. This precludes the application of Pyphy to many proteins, such as the histidine kinases and response regulators of two-component signal transduction systems, which show an enormous diversity in length and domain composition, but are nevertheless rewarding targets for phylogenetic analysis based on their conserved kinase and phospho-acceptor domains (14). For this reason, PhyloGenie contains a novel alignment routine, Blammer, which post-processes local pairwise sequence alignments obtained from BLAST or PSIBLAST (see Methods; Figure 2) to focus the resulting multiple alignment on conserved domains. Blammer extracts the BLAST HSPs above a given significant cut-off and coverage, converts them to a multiple alignment, identifies anchor regions of ungapped sequence and realigns the gapped regions in between using ClustalW. It then builds an HMM of the alignment and searches all full-length sequences that have BLAST HSPs in response to the original query for significant matches, which it realigns to obtain the final alignment. In addition, and unlike Pyphy, PhyloGenie allows users to customize all parameters in the search and alignment routines, thus making it possible to optimize PhyloGenie for specific questions.

Many approaches to tree inference exist and different methods may be used depending on the available computing infrastructure, the average size and the quality of alignments. By default, PhyloGenie provides a neighbor-joining (NJ) method (13), a fast and robust way to infer trees. This can be replaced by any program or script that accepts aligned FASTA format sequences and generates New Hampshire Bracket Format (Newick) trees. For example, PhyloGenie contains a script (treepuzzle.pl), which allows the use of Tree-Puzzle (15), one of the faster maximum likelihood tree inference programs. We believe that this solution is preferable to that implemented in Pyphy, which uses the program PAUP (16) for tree inference. PAUP is a proprietary program and uses a program-specific tree format.

A large repository of phylogenetic trees is of limited use unless a way of separating relevant from irrelevant trees for the question at hand is provided. For example, in evaluating the actinopterygian genome duplication hypothesis, Taylor *et al.* (17) examined large numbers of trees manually, as phylogenies proved difficult to analyze in an automated manner. To reduce the number of trees that have to be examined manually, PhyloGenie contains a tool that extracts phylogenies corresponding to specific, user-defined topological constraints from a database of trees. Pyphy circumvents this problem by focusing on a single phylogenetic hypothesis, namely LGT.

Application to a LGT hypothesis

Thermoplasma acidophilum is a thermoacidophilic euryarchaeon that lives at 59°C and pH 2, whose genome sequence suggests an extensive LGT with a phylogenetically distant organism, the crenarchaeote *S.solfataricus* that inhabits the same ecological niche (10). This transfer was deduced from the fact that 252 of 1478 genes of Thermoplasma had best BLAST matches to proteins of Sulfolobus. Since the Thermoplasma–Sulfolobus BLAST comparison was originally done before the completion of the Sulfolobus genome, we repeated it and obtained 303 genes for which best BLAST hits predicted a Sulfolobus sequence as the closest relative. A PhyloGenie search for LGTs between Thermoplasmata (*T.acidophilum*, *T.volcanium*, *Ferroplasma acidarmanus*, *Picrophilus torridus*)

and Sulfolobus (*S.solfataricus* and *S.tokodaii*) returned 185 trees. Of the 252 LGTs originally predicted from BLAST similarities (10), less than half were recovered by the PhyloGenie approach. An analysis with Pyphy returned 148 trees.

The potential LGTs are not distributed uniformly across the genome (Figure 4); the patterns of distribution are similar for the three methods, with local differences in the exact numbers. Globally, 93 LGTs were predicted by all three methods, 71 by PhyloGenie and BLAST, 40 by Pyphy and BLAST and 1 by PhyloGenie and Pyphy. A closer analysis as to why one method differed from the other two revealed that in the set of 40 proteins missed by PhyloGenie but predicted as LGTs by the other two methods, most were compatible with the lateral transfer hypothesis but were excluded due to low bootstrap support (Table 1). In the set of 71 proteins missed by Pyphy, 43 were due to the use of the reduced sequence database that Pyphy uses and the very stringent inclusion criteria for homologous sequences; this caused many alignments to miss relevant proteins and, in some cases, to consist solely of one protein from each of the two Thermoplasma species, *T.acidophilum* and *T.volcanium*. The one tree missed by BLAST is due to an archaeal sequence with a marginally better *E*-value than the closest Sulfolobus sequence relative.

In summary (Table 1): (i) 93 LGT predictions were supported by all three methods, and a further 90 were supported by at least two of the three methods and not contradicted by any; (ii) 8 LGTs were predicted by BLAST and PhyloGenie but contradicted by Pyphy, 13 were supported by BLAST and Pyphy but contradicted by PhyloGenie and 1 was supported by both Pyphy and PhyloGenie but contradicted by BLAST. Taking protein LGT predictions supported by at least two methods and not contradicted by any phylogenetic approach as true positives (184 trees), showed BLAST to be the most sensitive method with >99% sensitivity (183 of 184 true positives recovered) but also the least selective with 60% selectivity (303 predicted LGTs, 183 true positives). PhyloGenie

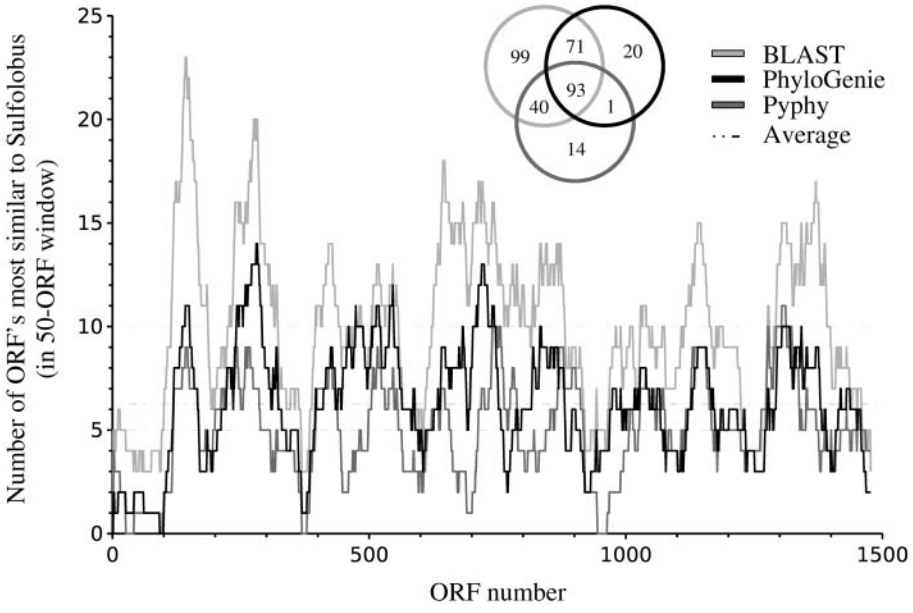


Figure 4. Chromosomal distribution of presumed laterally transferred ORFs between Thermoplasmata and Sulfolobus, according to PhyloGenie, Pyphy and best BLAST hits. The light gray, dark gray and black circles encompass the LGTs predicted by BLAST, Pyphy and PhyloGenie, respectively.

Table 1. Overview of LGT events identified by BLAST, Pyphy, and PhyloGenie

	LGT trees	Average bootstrap support	Negative trees[a]	Average bootstrap support	Thermoplasmata and Sulfolobus only (PhyloGenie/Pyphy)[b]	Trees missing Sulfolobus[c]	Additional sequences invalidate LGT[d]	Additional Archaea invalidate LGT[e]
BLAST	99	—	67	64 ± 30	—	—	—	—
PhyloGenie	20	79 ± 16	3	48 ± 4	1/0	Pyphy: 11	Pyphy: 2	Pyphy: 2
Pyphy	14	79 ± 20	8	87 ± 23	0/2	PhyloGenie: 3	PhyloGenie: 8	PhyloGenie: 7
BLAST + PhyloGenie	71	87 ± 15	8	67 ± 22	27/0	Pyphy: 43	Pyphy: 7	Pyphy: 1
BLAST + Pyphy	40	88 ± 16	13	62 ± 24	0/20	PhyloGenie: 1	PhyloGenie: 11	PhyloGenie: 7
PhyloGenie + Pyphy	1	79 ± 30	—	—	0/1	—	—	—
BLAST + PhyloGenie + Pyphy	93	90 ± 15	—	—	19/48	—	—	—

[a]The number of LGT predictions contradicted by a phylogeny based method.
[b]Trees based only on Thermoplasmata and Sulfolobus sequences and missing an outgroup.
[c]The number of trees for which the specified method failed to detect Sulfolobus homologs.
[d]The number of trees in which including additional sequences found by the specified method invalidate the other 's LGT prediction.
[e]The number of trees in which archaeal sequences invalidated the LGT prediction.

showed a sensitivity of 85% and selectivity of 85% whereas Pyphy had a sensitivity of 66% and a selectivity of 82%. Thus, of the three methods, PhyloGenie seems to be the one best able to combine high sensitivity with high selectivity. In detail, our conclusions are: (i) the Pyphy criteria for defining homologous sequences are too strict, thereby excluding many relevant sequences from analysis, as is apparent from the 43 true positives that were not predicted by Pyphy because it missed the Sulfolobus homologs (Table 1); (ii) less strict search criteria, as in PhyloGenie, circumvents this problem, but the resulting sequence diversity may lower bootstrap support in some cases to <50%, thereby causing trees supporting the LGT hypothesis to be missed, as in 24 of the 40 true positives missed by PhyloGenie; (iii) finally, as pointed out by Koski and Golding (2), best BLAST hits are of only moderate accuracy when identifying the phylogenetically nearest homolog.

The hypothesis of large-scale LGT between Thermoplasmata and Sulfolobus, proposed on the basis of best BLAST hits (10), is thus confirmed by phylogenetic analyses, albeit in a smaller number of cases than originally anticipated. The clustering of putatively transferred genes is also confirmed (Figure 4), pointing to a process that occurred mainly by the exchange of larger DNA regions.

Application to a genome duplication hypothesis

It has been proposed that the creation of metazoans and vertebrates from unicellular organisms would have been impossible without duplication of genes, as mechanisms evolving new functions at the price of discarding established ones would not provide an effective way to ''progress'' in evolution (18,19). Genome duplication was advanced as the primary source for new, redundant genes as it increases gene number without changing gene dosage. Dingerkus and Howell (11) proposed that the actinopterygian lineage (ray-finned fish), containing over 22 000 species, arose by means of tetraploidization due to the large number of chromosomes found in species whose ancestors diverged early in the evolution of actinopterygii. Support for this hypothesis has also been found in the seven hox clusters present in zebrafish (20), almost twice as many as in tetrapods, and in the 49 clades of orthologous proteins found by Taylor *et al*. (17).

An analysis of the set of trees derived by PhyloGenie from all zebrafish proteins present in the non-redundant GenBank database returned trees for 120 clades of orthologous genes, in which two *Danio rerio* proteins were present for one protein in *H. sapiens*, *M.musculus* and *G.gallus*. Of these, five had no discernible annotation information, 16 proteins seemed unlikely to be involved in development or gene regulation (synaptosome associated protein, chromobox4, photolyase, beta-carotene oxygenase, opsin I, Cytochrome P450, dystrophin, histocompatibility antigen class II, heat shock factor 1, heat shock factor 2, lamin1, lamin2, rhodopsin, troponin, and two subunits of a Na+/K+ transport ATPase) and the majority (99 proteins) consisted of morphogenic, growth factor and signal transduction proteins (33 HOX/PAX genes, 11 frizzled and other receptors, 9 FGF and other growth factors, various transcription factors, cyclases, kinases, etc.). In comparison, Pyphy returned 53 trees that matched the selection criteria. Of these, 8 had no discernible annotation, 11 seemed unlikely to be involved in development or regulation (laminin, axin, HSP, UQ-conjugator, etc.) and 34 were growth factors or signal transduction proteins (Frizzled, Hox, Pax, G-proteins, growth factors, etc.). The PhyloGenie analysis provides support for the genome duplication hypothesis advanced by Dingerkus and Howell (11) by more than doubling the number of supporting clades.

The analysis also suggests that a subsequent massive loss of non-morphogenic genes may have occurred. However, the *Danio rerio* and *G.gallus* genomes have not yet been completely determined. The large number of morphogenic and regulatory proteins we observe may therefore reflect, at least in part, an historic bias of molecular genetics towards development and cell cycle regulation. Support for this view comes from the observation that, if only two of the three tetrapod species are required, PhyloGenie and Pyphy return 351 and 118 trees, respectively, for nodes containing mouse and chicken, 331 and 141 trees for nodes with man and chicken, and 630 and 292 trees for nodes with man and mouse. It will only be possible to form an exact picture of the number and types of genes showing this 2:1 ratio, once completed genomes are available for a wide range of tetrapod and actinopterygian species.

CONCLUSIONS

We have introduced a new suite of programs for the generation and analysis of phylomes (PhyloGenie) and have compared its performance with that of a related software tool (Pyphy) on two previously studied phylogenetic problems. On attempting to detect LGTs in prokaryotic organisms, both methods seem to perform comparably. This ceases to be the case when analyses are attempted for which Pyphy was not designed, such as examining paralogous sequence relationships or sequence clades encompassing more than the immediate sequence relatives. In these cases, restrictive settings limit the ability of Pyphy to detect all relevant sequences. With regard to tree analysis, Pyphy is built to detect LGTs in a genome and graphically display the results. It does not support querying of more complex sequence relationships. In contrast, PhyloGenie is fully configurable in all parameters relating to sequence acquisition, alignment, and tree construction, and has the ability to filter the resulting database of trees for complex, user-defined tree topologies.

Automated methods are powerful, but also have drawbacks. The search and alignment parameters used for generating phylomes rely on assumptions and prior knowledge that may introduce errors or systematic biases. It is therefore essential to manually re-evaluate, for biological relevance, the steps and results between seed sequence and the phylogenies of interest. The problems encountered by PhyloGenie in the example analyses were mostly due to suboptimal search parameters that cause some alignments to contain large coiled-coil or low complexity regions, possibly convergently evolved features misleading phylogenetic inference, and splice isoforms or gene fragments complicating the automatic phylogenetic analysis. In addition, sampling bias, alignment errors, mutational saturation, long-branch attraction, methodological artifacts and differential gene loss can also account for the atypical placement of species in a tree. The results produced by PhyloGenie should therefore not be seen as the endpoint of

an analysis but rather as the first step in reducing the number of genes or alignments for which more time consuming, in depth analyses would need to be performed, before being able to draw conclusions with confidence.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank the referees for many helpful and constructive comments.

REFERENCES

1. Altschul,S.F., Madden,T.L., Schäffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
2. Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol*., **52**, 540–542.
3. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. Stanhope,M.J., Lupas,A.N., Italia,M.J., Koretke,K.K., Volker,C. and Brown,J.R. (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, **411**, 940–944.
5. Salzberg,S.L., White,O., Peterson,J. and Eisen,J.A. (2001) Microbial genes in the human genome: lateral transfer or gene loss? *Science*, **5523**, 1903–1906.
6. Roelofs,J. and Van Haastert,P.J. (2001) Genes lost during evolution. *Nature*, **411**, 1013–1014.
7. Sicheritz-Ponten,T. and Andersson,SG. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res*., **15**, 545–552.
8. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C.Phan,I., Pilbout,S. and Schneider,M. (2003) The Swiss-Prot protein knowledgebase and its supplement TREMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
9. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
10. Ruepp,A., Graml,W., Santos-Martinez,M.L., Koretke,K.K., Volker,C., Mewes,H.W., Frishman,D., Stocker,S., Lupas,A.N. and Baumeister,W. (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*, **407**, 508–513.
11. Dingerkus,G. and Howell,WM. (1976) Karyotypic analysis and evidence of tetraploidy in the North American paddlefish, *Polyodon spathula*. *Science*, **194**, 842–844.
12. Eddy,S.R. (1996) Hidden Markov Models. *Curr. Opin. Struct. Biol*., **6**, 361–365.
13. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., **4**, 406–425.
14. Koretke,K.K., Lupas,A.N., Warren,P.V., Rosenberg,M. and Brown,J.R. (2000) Evolution of two-component signal transduction. *Mol. Biol. Evol*., **17**, 1956–1970.
15. Schmidt,H.A., Strimmer,K., Vingron,M. and von Haeseler,A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.
16. Swofford,D.L. (1998) *PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods)*, version 4.0. Sinauer Associates, Sunderland, MA.
17. Taylor,J.S., Braasch,I., Frickey,T., Meyer,A. and Van de Peer,Y. (2003) Genome duplication, a trait shared by 22 000 species of ray-finned fish. *Genome Res*., **13**, 382–390.
18. Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, New York, NY.
19. Stephens,S.G. (1951) Possible significance of duplications in evolution. *Adv. Genet*., **4**, 247–265.
20. Amores,A., Force,A., Yan,Y.-L., Joly,L., Amemiya,C., Frity,A., Ho,R.K., Langeland,J., Prince,V., Wang,Y.-L., Westerfield,M., Ekker,M. and Postlethwait,J.H. (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science*, **282**, 1711–1714.

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 18 2004, pages 3702–3704
doi:10.1093/bioinformatics/bth444

CLANS: a Java application for visualizing protein families based on pairwise similarity

Tancred Frickey and Andrei Lupas*

Max Planck Institut fuer Entwicklungsbiologie, Spemannstrasse 35, 72076 Tuebingen, Germany

Received on April 8, 2004; revised on July 9, 2004; and accepted on July 26, 2004
Advance Access publication July 29, 2004

ABSTRACT

Summary: The main source of hypotheses on the structure and function of new proteins is their homology to proteins with known properties. Homologous relationships are typically established through sequence similarity searches, multiple alignments and phylogenetic reconstruction. In cases where the number of potential relationships is large, for example in P-loop NTPases with many thousands of members, alignments and phylogenies become computationally demanding, accumulate errors and lose resolution. In search of a better way to analyze relationships in large sequence datasets we have developed a Java application, CLANS (CLuster ANalysis of Sequences), which uses a version of the Fruchterman–Reingold graph layout algorithm to visualize pairwise sequence similarities in either two-dimensional or three-dimensional space.

Availability: CLANS can be downloaded at http://protevo.eb.tuebingen.mpg.de/download

Contact: andrei.lupas@tuebingen.mpg.de

INTRODUCTION

The use of homology to infer properties from known to previously unknown proteins is widespread throughout all domains of molecular biology. The most commonly used marker for homology is sequence similarity; this is so pervasive that sequence similarity searches are frequently, but inaccurately, referred to as homology searches. The standard tool for establishing sequence similarity is BLAST (Altschul *et al*., 1997). However, top-scoring BLAST matches do not necessarily identify the closest homologs (Koski and Golding, 2001). As an extreme example, BLAST identified 113 vertebrate genes with closest matches in bacteria, leading the International Human Genome Sequenceing Consortium, 2001 Human Genome Sequencing Consortium to propose multiple lateral transfer events from bacteria to vertebrates (2001). However, subsequent phylogenetic analysis was unable to support this hypothesis for any of the examined genes (Stanhope *et al*., 2001). The reasons top-scoring BLAST matches are used despite being clearly inferior to phylogenetic reconstruction are speed and ease of analysis. In a world of exponentially growing databases, situations are frequently encountered where phylogenetic reconstruction becomes computationally prohibitive. In addition, even before this point is reached, the prerequisite for multiple sequence alignments make phylogenetic inference cumbersome and prone to errors. This is in part because alignment complexity increases and accuracy decreases with the number of sequences, and in part because the limited number of phylogenetically informative sites in an alignment leads to loss of resolution.

An alternative approach that exploits the speed of BLAST and avoids the problems associated with multiple sequence alignments is the visualization of all-against-all pairwise similarities. This method can handle unrefined, unaligned data, including nonhomologous sequences. Unlike phylogenetic reconstruction it becomes more accurate with an increasing number of sequences, as the larger number of pairwise relationships average out the spurious matches that are the crux of simpler pairwise similarity-based analyses. BioLayout, a tool for visualizing such data based on the Fruchterman–Reingold (1991) graph layout algorithm, has previously been developed by Enright and Ouzounis (2001). Impediments in its use, including the prerequisite for precomputed similarities, limitations in changing the parameters for graph layout or the inability to add new sequences to existing graphs, prompted us to develop a new implementation.

IMPLEMENTATION

Similar to BioLayout (Enright and Ouzounis, 2001) we used a variant of the Fruchterman and Reingold graph layout algorithm to generate graphs providing a useful representation of pairwise sequence similarities. Sequences are represented by vertices in the graph, BLAST/PSIBLAST high scoring segment pairs (HSPs) are shown as edges connecting vertices and provide attractive forces proportional to the negative logarithm of the HSP's P-value. To keep all sequences from collapsing onto one point, a mild repulsive force is placed between all vertices. After random placement in either two-dimensional or three-dimensional space,

*To whom correspondence should be addressed.

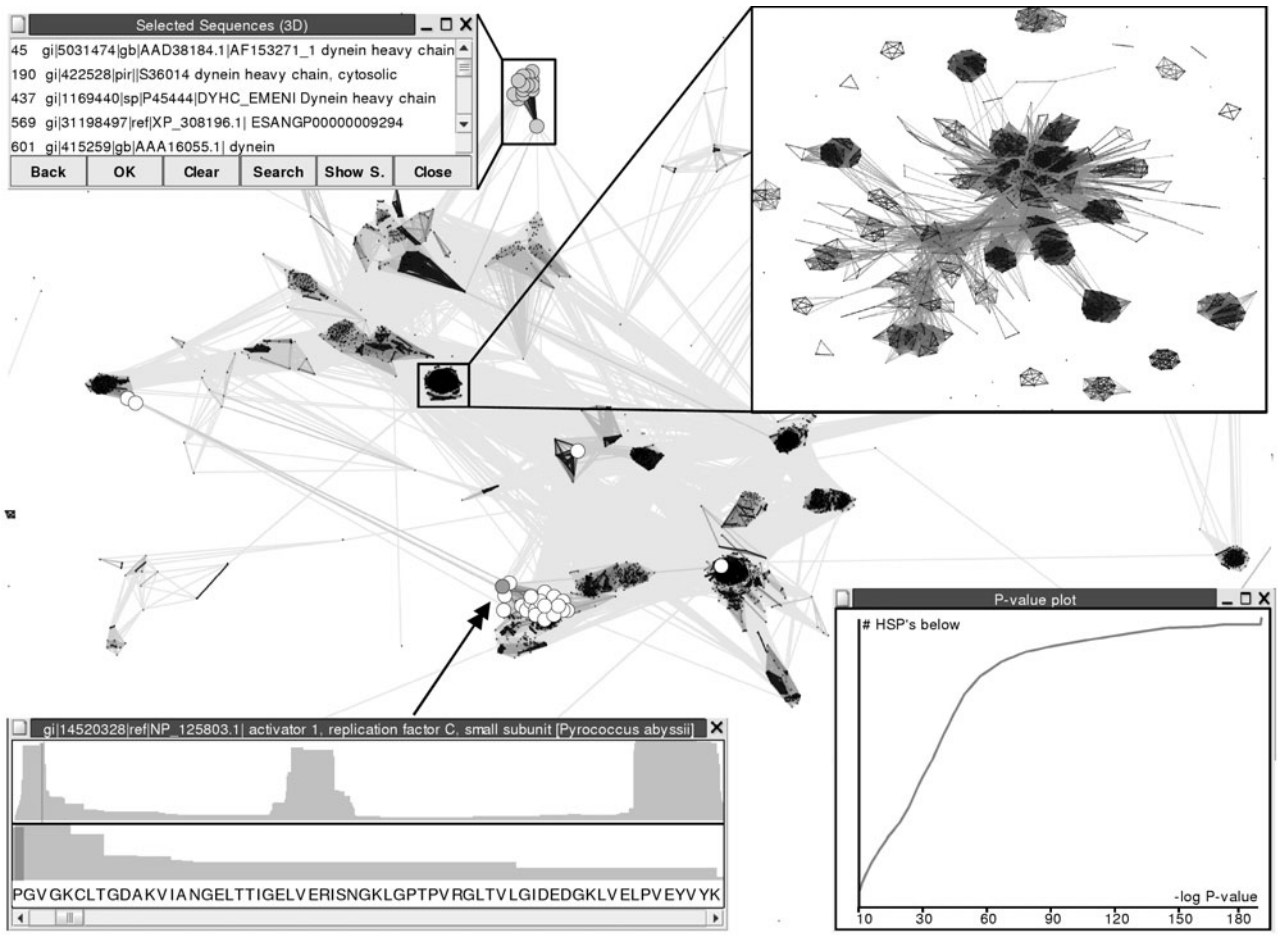


Fig. 1. Graph layout for 5101 AAA+-ATPases (Frickey and Lupas, 2004). Top right: a previously compact cluster (at P-values $\leq 10^{-10}$) disassembles and reveals a substructure when edges with P-values above 10^{-80} are removed. Top left: the names of currently selected sequences, highlighted by gray circles, are shown in a separate window. Botom right: graph of the P-value distribution of HSPs. Bottom left: a sequence is selected from the dataset and used as BLAST query (double arrows). All HSPs are collected, mapped onto the sequence and displayed in a separate window. The top panel shows the distribution of HSPs over the query, the bottom panel a closeup with individual residues visible. When selecting a residue, for example Proline in the N-terminal region of the sequence (gray vertical bar), all sequences with HSPs covering that residue are highlighted by white circles in the main graph. The tripartite distribution of BLAST hits for the query is notable. Only the most N-terminal region of the query provides HSPs that connect it to sequences outside its cluster; more C-terminal residues have BLAST matches only within the cluster.

the vertices are moved iteratively according to the force vectors resulting from all pairwise interactions until the overall vertex movement becomes negligible. While this approach, coupled with random placement, causes non-deterministic behavior, similar sequences or sequence groups reproducibly come to lie close together after a few iterations thus generating similar, although non-identical graphs for different runs.

APPLICATION

Default input is a file with protein sequences in FASTA format. Command line parameters specify the location of BLAST/PSIBLAST executables, databases and search options. The program performs all-against-all BLAST searches and calculates pairwise attraction values based on the HSP P-values. The graph showing all pairwise interactions can be rotated, translated and zoomed to better view sequence relationships. Discarding P-values above a certain cutoff and varying that value can cause previously compact groups to disaggregate and reveal their substructure (Fig. 1). The user can add, extract or remove sequence from graphs, use varying P-value cutoffs for layout of different graph regions and, for comparative purposes, infer NJ-trees based on either the BLAST P-values or distances separating vertices in the graph. For increased sensitivity, necessary when distantly related groups of proteins are to be viewed, it is possible to evolve PSIBLAST profiles over a reference database and

subsequently use these profiles to collect the HSPs used for graph layout.

As an alternative to FASTA input, it is possible to load a matrix of precomputed attraction values and thereby display any kind of data based on pairwise interactions. Examples might be visualization of social networks to determine key organisms most likely to rapidly disseminate diseases or layout of bacterial species based on similarities in metabolism and lifestyle (trophies, antibiotic sensitivities, etc.).

A Java1.4 or higher runtime environment is necessary to run CLANS and installation of BLAST/PSIBLAST is required for full functionality.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3444.

Enright,A.J. and Ouzounis,C.A. (2001) BioLayout-an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, **17**, 853–854.

Frickey,T. and Lupas,A.N. (2004) Phylogenetic analysis of AAA proteins. *J. Struct. Biol.*, **146**, 2–10.

Fruchterman,T.M. and Reingold,E.M., (1991) Force directed placement. *Softw. Pract. Exp.*, **21**, 1129–1164.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.

International Human Genome Sequenceing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Stanhope,M.J., Lupas,A.N., Italia,M.J., Koretke,K.K., Volker,C. and Brown,J.R. (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, **411**, 940–944.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Structural Biology 146 (2004) 2–10
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

Phylogenetic analysis of AAA proteins

Tancred Frickey and Andrei N. Lupas*

Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Spemannstr. 35, Tübingen D-72076, Germany

Received 18 September 2003, and in revised form 31 October 2003

Abstract

AAA ATPases form a large protein family with manifold cellular roles. They belong to the AAA+ superfamily of ringshaped P-loop NTPases, which exert their activity through the energy-dependent unfolding of macromolecules. Phylogenetic analyses have suggested the existence of five major clades of AAA domains (proteasome subunits, metalloproteases, domains D1 and D2 of ATPases with two AAA domains, and the MSP1/katanin/spastin group), as well as a number of deeply branching minor clades. These analyses however have been characterized by a lack of consistency in defining the boundaries of the AAA family. We have used cluster analysis to delineate unambiguously the group of AAA sequences within the AAA+ superfamily. Phylogenetic and cluster analysis of this sequence set revealed the existence of a sixth major AAA clade, comprising the mitochondrial, membrane-bound protein BCS1 and its homologues. In addition, we identified several deep branches consisting mainly of hypothetical proteins resulting from genomic projects. Analysis of the AAA N-domains provided direct support for the obtained phylogeny for most branches, but revealed some deep splits that had not been apparent from phylogenetic analysis and some unexpected similarities between distant clades. It also revealed highly degenerate D1 domains in plant MSP1 sequences and in at least one deeply branching group of hypothetical proteins (YC46), showing that AAA proteins with two ATPase domains arose at least three times independently.

Keywords: AAA proteins; ATPases; Classification; Cluster analysis; Phylogeny

1. Introduction

AAA proteins were first described by Erdmann et al. (1991) as a new family of 'ATPases Associated with diverse cellular Activities.' The family is characterized by a highly conserved P-loop NTPase domain of about 240 residues, which, in addition to the hallmark Walker A and B motifs, contains further regions of high sequence conservation, most conspicuously the so-called 'second region of homology' (SRH) (Tomoyasu et al., 1993). All AAA proteins whose oligomeric structure has been investigated form hexameric rings, although in some cases, such as that of katanin, oligomerization may only occur under certain conditions (Hartman and Vale, 1999). The domain architecture of AAA proteins consists of a non-ATPase, N-terminal domain (the N-domain), considered to be the primary substrate recognition site, followed by either one or two AAA domains (named D1 and D2). In proteins with two AAA domains one domain may be degenerate, such as for example D1 in peroxisomal ATPases and D2 in Sec18/NSF. Functionally, AAA proteins have been implicated in protein degradation, maturation of membrane complexes, gene expression, homo- and heterotypic membrane fusion, and microtubule disassembly. Mechanistically, they are thought to exert their activity through the energy-dependent disassembly and unfolding of proteins. Several crystal structures of AAA proteins have been determined, most recently the complete structure of p97, an ATPase with two canonical AAA domains (DeLaBarre and Brunger, 2003). These structures have shown that the SRH is located away from the nucleotide-binding pocket of the ATPase domain, such that, in a ring-shaped arrangement, the SRH of one subunit projects an arginine residue (the 'arginine finger') into the nucleotide-binding pocket of the next subunit in the ring.

* Corresponding author. Fax: +49-7071-601-349.
E-mail address: andrei.lupas@tuebingen.mpg.de (A.N. Lupas).

1047-8477/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2003.11.020

This observation has suggested a mechanism for concerted nucleotide hydrolysis and provides an explanation for the high degree of sequence conservation in the SRH (Lupas and Martin, 2002).

AAA proteins are a large and diverse family and their phylogeny has been analysed repeatedly over the years (Frohlich, 2001; Beyer, 1997; Swaffield and Purugganan, 1997; Wolf et al., 1998; see also http://aaa-proteins.uni-graz.at/AAA/Tree.html). These analyses varied in their approach, in the sequences included, and in the treatment of proteins with two AAA domains. Nevertheless, a reasonably consistent picture emerged of five main clades of AAA domains, corresponding to D1, D2, proteasome subunits, metalloproteases, and to a loosely defined 'meiotic' group comprising katanins, spastins, and MSP1. Some details of the trees remained puzzling, for example the fairly consistent grouping of Sec18/NSF D1 in the D2 clade. However, the most important shortcoming of these analyses was the inconsistent way in which sequences were selected: On the one hand, sequences that contained all canonical residues (Walker A and B, sensor-1, SRH) were sometimes excluded as too divergent; on the other hand, clearly degenerate sequences (usually corresponding to the inactive domains of ATPases with two AAA domains) were included, even though it is well known that inactive sequences evolve at a much higher rate and therefore confuse the deep branching order in phylogenetic analyses (see for example our discussion of the branching order for subunits of the 20S proteasome and the 11S regulator (Volker and Lupas, 2002)). Last year, we proposed a classification of AAA proteins within the AAA+ superfamily, based on the presence of the SRH (Lupas and Martin, 2002). Here, we used an alternative, automated approach (cluster analysis) to delineate unambiguously the AAA family. Analysis of this sequence set allowed us to derive a comprehensive picture of the phylogenetic relationships in currently known AAA proteins.

2. Methods

2.1. Selection of AAA+ proteins

In a first pass, the non-redundant protein sequence database (nr) at the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov) was searched by seeding PSI-BLAST (Altschul et al., 1997) with the AAA domain alignment of the SMART database (smart.embl-heidelberg.de; Schultz et al., 1998). All sequences with expect-values (*E*-values) of 10,000 or less were extracted and collected into a new database, which was searched with a Hidden Markov Model (HMM) of the AAA+ domain, using HMMer (hmmer.wustl.edu). The HMM was derived from the alphaA to alphaE region of the alignment of AAA+ domains published by Lupas et al. (1997), calibrated with 20,000 samples. Sequences that matched the HMM at *E*-values of 1 or less were extracted and formed our extended AAA+ set. The parameters were chosen such that sequences related to AAA+ but clearly outside this set were included, so as to ensure that we had obtained a comprehensive representation of the AAA+ family.

2.2. Cluster analysis of AAA+ proteins and selection of AAA sequences

Clustering of the resulting dataset was performed using a JAVA program, CLANS, which is based on the Fruchterman–Reingold algorithm (1991) and which we developed as part of a project designed to evaluate the accurracy of multiple alignments. The program resembles BioLayout (Enright and Ouzounis, 2001) and will be described in detail elsewhere; briefly, it uses the probability-values (*P*-values) of high-scoring segment pairs (HSPs) obtained from an $N \times N$ BLAST search, to compute attractive and repulsive forces between each sequence pair. A three-dimensional representation is achieved by randomly seeding sequences in space. The sequences are then moved within this environment according to the force vectors resulting from all pairwise interactions and the process is repeated to convergence. Clustering of the AAA+ set using BLAST *P*-values of 10^{-10} or less yielded a compact, well-defined cluster of AAA proteins, which were extracted for further analysis.

2.3. Selection and analysis of AAA and N-domains

A representative HMM of the AAA domain was derived from regions alphaA to alphaC of the AAA sequences in the alignment of Lupas et al. (1997) and calibrated with 20,000 samples. This HMM was used to identify AAA domains within the set of AAA proteins, using HMMer at an *E*-value cutoff of 10. The identified AAA domains and all sequences C-terminal to them were masked out to obtain the set of N-domains. CLANS was used subsequently to cluster AAA domains at BLAST *P*-values below 10^{-30} and N-domains at PSI-BLAST *P*-values below 10^{-3} (after five iterations).

After exclusion of a small number of visibly degenerate domains that had satisfied the relaxed cutoff chosen for the HMMer search, insert regions present in less than 1% of the sequences were removed. AAA domains were then aligned in ClustalW (Thompson et al., 1994) and gapped regions were adjusted so as to maintain the integrity of core secondary structure elements. The resulting manually refined alignment was used for phylogenetic tree construction. Phylogenies were inferred using the Asatura software (Van de Peer et al, 2002) in combination with the jtt (Jones et al., 1992) and mtrev (Adachi and Hasegawa, 1996) substitution matrices. Due to the varying rates of evolution in the different

branches, trees and subtrees calculated under different saturation correction conditions were later combined to yield one tree, which best reflects the branching pattern of the individual branches.

3. Cluster analysis of AAA+ proteins and definition of the AAA family

The terms 'AAA' and 'AAA+' are often used interchangeably, even though 'AAA' refers to a subset of proteins within 'AAA+.' For example, the SMART and Pfam databases labeled their AAA+ HMMs as 'AAA.' In fact, neither term is well-defined. For this reason, the set of sequences included in phylogenetic analyses of the AAA family has been quite variable and all studies have pointed to the existence of a substantial number of sequences that are difficult to assign (Beyer, 1997; Frohlich, 2001; Swaffield and Purugganan, 1997). We believe that AAA proteins are actually clearly distinguishable from related sequences and recently proposed a set of morphological characteristics to define AAA and AAA+ proteins in relation to each other and to NTPases of the RecA fold (Lupas and Martin, 2002). These characteristics, however, are difficult to use in an automated fashion, posing problems for the analysis of the non-redundant (nr) database and thus also for a comprehensive phylogeny of AAA proteins. We therefore used cluster analysis of *P*-values from pairwise sequence comparisons to investigate whether AAA proteins could be identified unambiguously using a scalable, automated method.

A PSI-BLAST search of the nr database at NCBI using the SMART AAA alignment as a seed yielded 49,966 sequences at an *E*-value cutoff of 10,000. These sequences were then searched with an HMM based on a manually curated alignment (see Section 2) to identify AAA+ proteins as well as some of their nearest neighbors in sequence space, 5101 sequences in all. Although we obtained a fairly comprehensive representation of AAA+ proteins, the primary purpose of these successive steps was to focus on the AAA+ group while not excluding any sequence of the AAA subset, rather than to achieve a complete enumeration of AAA+ sequences. The set of 5101 sequences was subjected to cluster analysis and revealed a compact cluster for AAA proteins (Fig. 1A), which was well separated from neighboring clusters. These sequences, 1241 in total, formed our AAA set.

Although the purpose of the clustering step was not to provide a comprehensive picture of AAA+, most of the important sequence families should be represented in Fig. 1A. A core group of 23 clusters is strongly interconnected by pairwise relationships and includes all classical AAA+ proteins, including Clp/Hsp100, Lon, AAA, RuvB, Mg chelatase, clamp loaders, and σ54-dependent transcription factors (NtrC). This core group also contains three clusters of open reading frames (ORFs) that are as yet unannotated. One may further observe the presence of superclusters within the core group, such as for example the one formed of clusters 26–31, but we would caution against interpreting these as phylogenetically relevant at this level of the analysis. Nevertheless, the grouping for example of replication factor C, clamp loader subunits, Werner helicase interacting protein and RuvB (clusters 19–22) seems biologically reasonable and warrants further investigation. We obtained seven clusters that were peripheral but connected to the core group; these included DnaA, McrB, and dynein, which have been previously associated with the AAA+ superfamily. Finally, some clusters that are known to form an outgroup to AAA+, such as that of ABC transporters, remained unconnected.

4. Phylogenetic analysis of AAA proteins

We extracted all AAA domains from the AAA set of 1241 sequences, using HMM searches based on a manually curated seed alignment at a very relaxed *E*-value of 10 (see Section 2). The seed alignment included the AAA domain sequence from the N-terminal α-helix to the first helix of the C-terminal helical extension. We obained 1369 domain sequences, of which we subsequently excluded 81 because they lacked at least one (and generally several) of the canonical residues: GKT/S in the Walker A region, D/E in the Walker B region, a hydrogen-donor residue in the sensor-1 region, and the arginine in the SRH. The excluded sequences are most likely catalytically inactive and therefore prone to more rapid evolutionary divergence, leading to problems in phylogenetic reconstruction (Gribaldo and Philippe, 2002). We generated a manually curated alignment from the remaining 1288 domain sequences and subjected it to distance-based phylogenetic analysis. Due to the selective constraints on the domain (seen in the high degree of sequence conservation) and the long divergence time (seen in the broad phyletic representation), we expected the alignment to show a high degree of mutational saturation. Correction for this saturation reduces the resolution of closely related clades, but provides a better representation of the deep branching order due to reduction of homoplasy (Van de Peer et al, 2002). For this reason, we decided to compute the overall tree using stringent saturation correction (Fig. 2A) and then re-evaluate the major clades independently. The resulting composite tree (Fig. 2B) yielded six major clades, as well as several smaller, long-branching, deeply rooted minor clades, some of which changed their positions substantially upon minor changes in procedure.

An open question of this analysis regarded the position of the root. The traditional way of rooting a tree is inclusion of an outgroup in the analysis, but in this case

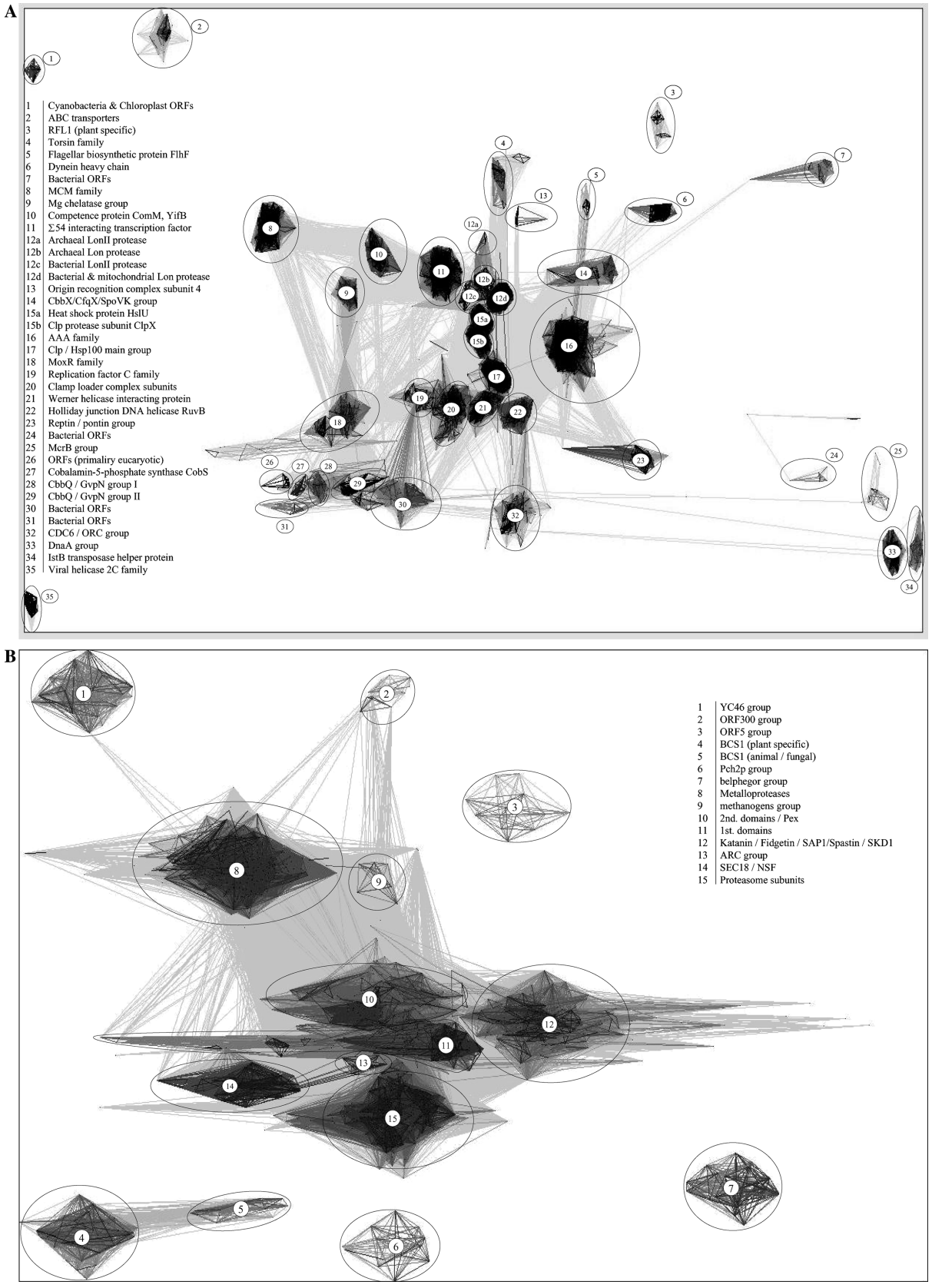


Fig. 1. (A) Cluster analysis of full-length sequences for the extended AAA+ dataset. Line coloring reflects BLAST *P*-values; dark lines represent pairwise connections with very low, lighter lines those with *P*-values closer to the cutoff (10^{-10}). Cluster 16 comprises the AAA sequences. (B) Cluster analysis of AAA domains at a cutoff *P*-value of 10^{-30}. The clades of the AAA tree are recovered as separate clusters.

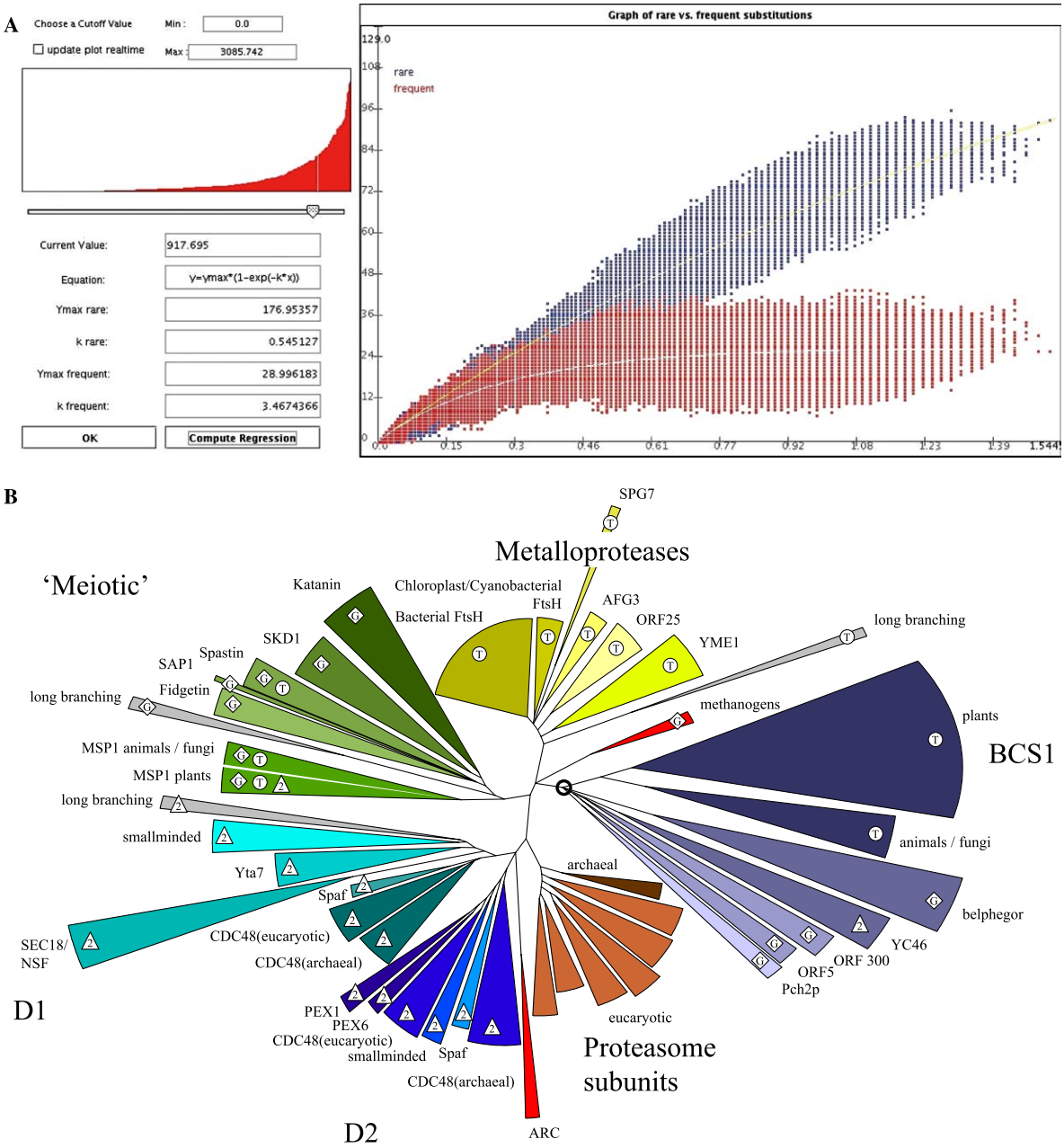


Fig. 2. (A) Saturation correction plot for the determination of the deep branching order in the AAA tree. The number of amino acid changes for the 'rare' and 'frequent' datasets are plotted against pairwise distance. Clear mutational saturation, where increased evolutionary distance is not reflected in a higher number of amino acid changes, is observable for the 'frequent' subset. (B) Phylogeny of AAA domains. The depicted phylogeny is a composite tree, formed of subtrees calculated under various saturation correction scenarios. Length and angle of each clade reflect the maximum branch length and number of sequences belonging to that clade. The presumed root of the tree is marked by a black circle, long branching minor clades whose monophyly is unsure are colored in gray. The letter "T" indicates clades containing transmembrane helices in their N-domains, "G" denotes clades containing a gapped SRH and "2" specifies clades of sequences containing two ATPase domains, either canonical or degenerate.

it is difficult to identify the appropriate outgroup and the alignment is already mutationally saturated, so that adding more distantly related sequences would only make the analysis of the deep branching order more unreliable. As an alternative approach, we did a cluster analysis of AAA domains and observed the emergence of pairwise connections between clusters at different cutoffs. The analysis yielded a core group of clusters, comprising the five previously defined main clades of AAA domains, as well as a number of peripheral clusters (Fig. 1B). As the cutoff was made less stringent, connections appeared between the peripheral and core groups; these connections were radial into the core group, rather than circular between the peripheral groups, with two exceptions: the ORF300 and ORF5 clusters connected together, as did the plant and animal

BCS1 clusters. This observation suggests to us a position of the root close to the origin of the peripheral groups, as shown in Fig. 2B.

4.1. Major clades

There are six major clades in the AAA tree: metalloproteases (in bacteria and organelles), 'meiotic' proteins (in eukaryotes and a small number of archaea), D1 and D2 domains of proteins with two AAA domains (in archaea and eukaryotes, as well as in a small number of bacteria, presumably by lateral transfer), proteasome subunits (in archaea and eukaryotes; note that ARC has a similar N-domain and may be an extremely divergent member of this clade in actinomycetes), and BCS1 (in eukaryotes).

4.1.1. Metalloproteases

Proteins in this group have a specific domain structure, with a trans-membrane N-domain, a central AAA domain and a C-terminal metalloprotease domain. They are found in the inner membranes of bacteria, where they degrade both soluble and membrane proteins, and organelles, where they primarily degrade unassembled subunits of membrane complexes. Their branching pattern reflects functional groups and holds few surprises, except maybe for the occurrence of a deep branch of sequences specific for cyanobacteria and chloroplasts (named for ORF25 of *Porphyra purpurea*), whose functions are unknown. Branch lengths are among the shortest in the AAA tree, with two exceptions: SPG7/paraplegin and a basal group of sequences that are probably driven together by long-branch attraction and for the most part lack the active-site residues of the protease domain. Cluster analysis of the N-domains (Fig. 3) groups together all metalloproteases except YME1 and reveals a deep split between plant YME1 on one hand and animal and fungal YME1 on the other, whose N-domains share no apparent sequence similarity. The depth of this split was not expected from their separation in the tree.

4.1.2. Meiotic group

The main branches in this group are katanin (MEI-1), SKD1 (Vps4), spastin, fidgetin, and MSP1. The functions of 'meiotic' proteins are poorly characterized at the molecular level and it is therefore unclear what properties distinguish this group from other AAA proteins. SKD1 is involved in vacuolar sorting and endosomal transport, katanin, and MEI-1 are microtubule severing proteins implicated in mitosis and meiosis, and spastins are involved in microtubule disassembly. Features of this

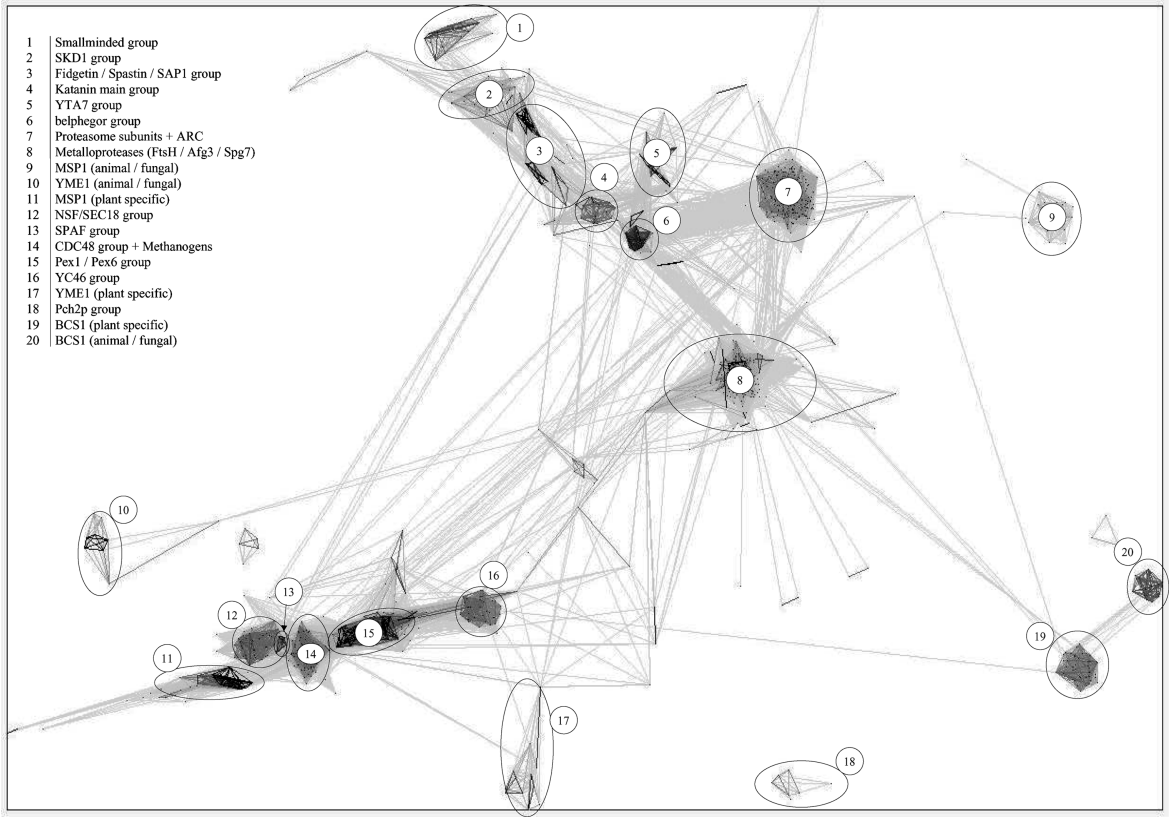


Fig. 3. Cluster analysis of N-domains. Sequences were clustered using PSI-BLAST *P*-values of 10^{-3} as a cutoff.

clade that are consistently recovered in different analyses are the basal position of MSP1 and the monophyly of spastin, fidgetin, and SAP1. In addition, the group contains a longer, deep branch formed mainly by archaeal sequences; the monophyly of this group is unclear. A unifying feature of 'meiotic' proteins is a gap of two residues in the SRH, immediately preceeding the arginine finger; all sequences in this group have the gap (except for *Thermoplasma* Ta1175 and its closest homologs). We proposed earlier that this might be a morphological trait specific to this group (Lupas and Martin, 2002), but the current analysis shows that several minor clades (methanogens, belphegor, ORF300, ORF5, and Pch2p), mostly branching at the root of the tree, also contain this gap. The most surprising finding in this group resulted from the analysis of the N-domains, which showed that the large N-domain found in plant MSP1 sequences, but not in those from animals or fungi, actually contained a degenerate ATPase domain, revealing a duplication event independent of the one leading to the main group of proteins with two AAA domains (Cdc48/p97).

4.1.3. D1 and D2 domains

The originally defined AAA proteins with two ATPase domains consisted of Cdc48/p97, Sec18/NSF, Pex1 and Pex6, YTA7, SPAF, and smallminded. These include some of the best-characterized AAA proteins and seem to be primarily involved in membrane fusion processes. Cdc48 has also been implicated in proteasome-dependent protein degradation. The D1 and D2 domains from these proteins were recovered consistently in their respective clades: NSF and YTA7, which contain a degenerate D2 domain, were found in the D1 clade and PEX1 and PEX6, which contain a degenerate D1 domain, were found in the D2 clade; previously, most analyses had recovered NSF D1 in the D2 clade. It was previously suggested that all ATPases with two AAA domains originated from a single ancestor, based on the homology of their N-domains (Golbik et al., 1999). Our cluster analysis of N-domains supports monophyly of the originally defined group (but note that no similarity was found between the N-domains of smallminded and YTA7 and the other N-domains); however, we discovered at least two other, independent duplication events in the AAA tree. Two proteins in this clade were predicted to contain transmembrane segments: a Cdc48 homolog from *Plasmodium falciparum* (gi23612492), which contains a unique domain with a transmembrane segment N-terminal to the conserved N-domain of Cdc48 proteins, and a deeply branching protein from *Arabidopsis thaliana* (gi15227690), which appears to have no N-domain and contains the transmembrane segment at its C-terminus (confirmed by EST matches to *Arabidopsis* and *Beta vulgaris*).

4.1.4. 26S proteasome subunits

Proteins in this clade form homo-oligomeric complexes in archaea and hetero-oligomeric complexes in eukaryotes; they activate proteolysis by opening the central channel of the 20S proteasome (which gives access to the proteolytic chamber) and by unfolding and translocating substrates. This clade showed little mutational saturation and consequently held few surprises. The previously described and well-supported phylogenetic pattern, including the basal position of archeal subunits and monophyly of the 6 eucaryotic subunits, was consistently recovered. Our analysis of N-domains yielded a well-delineated cluster for this group, which, in addition to the proteasomal subunits also contained ARC. We investigated whether this might be due to the coiled-coil segment present in ARC and in proteasome subunits, but we found that significant similarity extended to the region between the coiled coil and the AAA domain, which is rich in β-strands and may form a β-barrel. Surprisingly, two proteasome ATPase subunits from *Plasmodium yoellii yoellii* were predicted to contain transmembrane segments: the PRS7 homolog (gi23490548) two segments (one N- and one C-terminal), and the PRS8 homolog (gi23490696) one segment (C-terminal). Structurally, an attachment of the proteasome to the membrane via these segments seems possible; however, the biological significance is unclear, as these segments are not found in the homologs of even the closest relative, *P. falciparum*.

4.1.5. BCS1

BCS1 is a protein of the mitochondrial inner membrane (Folsch et al., 1996), which appears to chaperone the assembly of membrane complexes. In yeast, it is required for the formation of functional Rieske iron–sulfur protein (Nobrega et al., 1992). The BCS1 clade shows a deep division between plant sequences and animal and fungal sequences. Of interest hereby is the high number of paralogs present in plants (36 paralogs for *A. thaliana*) compared with other eukaryotes (2 genes per fungus, 1 gene per animal). N-terminal sequence clustering recovered the deep division between these two groups with only a few, tentative connections between them. One ORF (gi25411838 of *Arabidopsis*) contained two nearly identical copies of the BCS1 protein, but we could not confirm it with EST searches. Should this ORF prove to be a protein, it would represent a very recent addition to the group of ATPases with two AAA domains.

4.2. Minor clades

The minor clades are mostly small groups of long-branching sequences, located at the base of the major clades. They are ARC, basal to D1 and D2; a methanogen group, basal to metalloproteases; and five groups (belphegor, Pch2p, ORF300, ORF5, and YC46)

radiating from the presumed root of the tree. In addition, two proteins of Sulfolobus (*Sulfolobus solfataricus* Sso2831 and Sso2420 and their homologs) could not be assigned to any branch and varied widely in phylogenetic position between analyses; they are therefore not shown in Fig. 2B. These two proteins have very short N-domains, consisting primarily of a transmembrane helix, and contain two well-conserved AAA domains. They may thus represent a further independent duplication event in the AAA tree.

ARC, long-branching and specific to actinomycetes, appears basal to D1 and D2 in the tree but resembles proteasomal subunits in the N-domain. ARC is encoded in chromosomal vicinity to the operon of 20S proteasome subunits and has been suspected since its discovery to be the bacterial equivalent of 26S ATPases, but an interaction or activation of the proteasome with ARC could not be substantiated experimentally (Wolf et al., 1998).

The 'methanogens' group, purely archaeal and consistently recovered as monophyletic, variously occurred either basally within the metalloprotease clade or close to the presumed root of the tree. There is no reason however to presume a closer relationship to metalloproteases, as proteins in this group do not contain a protease domain and have a gapped SRH sequence. The group contains Methanococcus Mj1494, Archaeoglobus Af1285, Methanobacterium Mth1011, Methanopyrus Mk1368, Methanosarcina Mm0304 (and homologs). Cluster analysis grouped the N-domains of the methanogens with Cdc48 and sequence analysis by PSI-BLAST and 3D-PSSM (Kelley et al., 2000; www.sbg.bio.ic.ac.uk/~3dpssm) showed that they are likely to assume a β-clam structure homologous to that found in the N-domains of the Cdc48 group (Coles et al., 1999).

Five minor clades, Pch2P (eukaryotes), ORF5 (bacteria; named for *Photorhabdus luminescens* ORF5), ORF300 (bacteria and archaea; named for *Escherichia coli* ORF300), belphegor (eukaryotes) and YC46 (cyanobacteria and red algal chloroplasts; named for YC46 of *Odontella sinensis*), radiate from the presumed root of the tree; their branching pattern is unclear. Common to these clades are longer than average branch lengths and a basal position to BCS1 in most phylogenies. The ORF5 and ORF300 clades were monophyletic in some trees and ORF5 had its closest connections to ORF300 in cluster analysis, suggesting a closer evolutionary relationship between these clades. Analysis of the N-domains surprisingly revealed the existence of a degenerate ATPase domain in the YC46 clade.

Four of the five minor clades originating near the root have a gapped SRH and alignments to AAA+ proteins in the region of the 'arginine finger' point to this being an ancestral trait. We therefore favor the scenario that the ancestor of YC46 acquired an insertion of two residues, to obtain the RPGR consensus sequence, subsequently forming the main AAA clades observed today. The only exception among the main clades is the 'meiotic' group, whose SRH is gapped. We may have incorrectly reconstructed its position relative to the root; however, in view of its strong clustering with the other main AAA clades (Fig. 1B), we think that the 'meiotic' ancestor may have reverted to a gapped SRH. Such spontaneous reversions are observable in several proteins, whose nearest homologs have a canonical SRH (see for example Neurospora crassa B1K11.010 versus its homologs in *Schizosaccharomyces* and *Saccharomyces*, or *Drosophila* CG12010-PA versus its homologs in Anopheles and human).

5. Conclusions

In this paper, we used cluster analysis to outline the AAA sequences within the AAA+ superfamily and subjected them to phylogenetic analysis. Our approach differs from the ones previously taken by the completeness and consistency of the sequence dataset and by the use of a correction procedure for mutational saturation. Our analysis recovered the five major, well-accepted clades of AAA proteins, consisting of proteasome subunits, metalloproteases, domains D1 and D2 of ATPases with two AAA domains, and the MSP1/katanin/spastin group, as well as a sixth one, consisting of BCS1 and its homologs. In addition, we identified a number of minor clades, most of them novel and for the most part branching close to the presumed root of the tree, which we located tentatively using cluster analysis of the AAA sequence set. Most of the minor clades were prokaryotic, in contrast to the well-established AAA clades, which are primarily eukaryotic. This suggests that the AAA family was already diverse prior to the separation of the three domains of life.

Our most surprising findings concerned: (I) the extreme paralogy of plant BCS1, with 36 copies in *Arabidopsis* alone, (II) the polyphyletic origin of AAA proteins with two ATPase domains, (III) the N-domain homologies between the Cdc48/p97 group and a distantly branching group of proteins from archaeal methanogens, as well as between ARC and proteasomal ATPases, (IV) the dissimilarity of N-domains in some groups, such as for example between MSP1 and other 'meiotic' proteins, or between plant YME1, animal and fungal YME1, and all other metalloproteases, and (V) the ancestral and probably polyphyletic nature of a variant second region of homology, which lacks two residues preceding the 'arginine finger.'

We believe that our AAA tree is the most reliable yet computed. However, we believe it to be only as complete as the current sequence database. The multiple, deep-branching clades formed primarily of genomic ORFs

suggest that further AAA clades will emerge in future genome projects and that some currently minor clades will be recognized as major clades. In addition, some small, basal, long-branching sequence groups that are currently placed within one of the major clades may well expand to form their own separate clades, such as for example the already mentioned membrane-associated *Arabidopsis* ORF (gi15227690), which differs substantially from other D1/D2 proteins and is not grouped with any major clade in cluster analysis.

Correlating functional information with the individual clades showed us how little is known as yet about this diverse and important protein family.

References

Adachi, J., Hasegawa, M., 1996. Abstract model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42 (4), 459–468.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Beyer, A., 1997. Abstract sequence analysis of the AAA protein family. Protein Sci. 6 (10), 2043–2058.

Coles, M., Diercks, T., Liermann, J., Groger, A., Rockel, B., Baumeister, W., Koretke, K.K., Lupas, A., Peters, J., Kessler, H., 1999. The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple beta-alphabetabeta element. Curr. Biol. 9 (20), 1158–1168.

DeLaBarre, B., Brunger, A.T., 2003. Complete structure of p97/valosin-containing protein reveals communication between nucleotide domains. Nat. Struct. Biol. 10, 856–863.

Enright, A.J., Ouzounis, C.A., 2001. BioLayout—an automatic graph layout algorithm for similarity visualization. Bioinformatics 17, 853–854.

Erdmann, R., Wiebel, F.F., Flessau, A., Rytka, J., Beyer, A., Frohlich, K.U., Kunau, W.H., 1991. PAS1, a yeast gene required for peroxisome biogenesis, encodes a member of a novel family of putative ATPases. Cell 64 (3), 499–510.

Folsch, H., Guiard, B., Neupert, W., Stuart, R.A., 1996. Internal targeting signal of the BCS1 protein: a novel mechanism of import into mitochondria. EMBO J. 15 (3), 479–487.

Frohlich, K.U., 2001. An AAA family tree. J. Cell Sci. 114 (Pt 9), 1601–1602.

Fruchterman, T.M., Reingold, E.M., 1991. Graph drawing by force-directed placement. Software Pract. Exp. 21, 1129–1164.

Golbik, R., Lupas, A.N., Koretke, K.K., Baumeister, W., Peters, J., 1999. The Janus face of the archaeal Cdc48/p97 homologue VAT: protein folding versus unfolding. Biol. Chem. 380 (9), 1049–1062.

Gribaldo, S., Philippe, H., 2002. Ancient phylogenetic relationships. Theor. Popul. Biol. 61 (4), 391–408.

Hartman, J.J., Vale, R.D., 1999. Microtubule disassembly by ATP-dependent oligomerization of the AAA enzyme katanin. Science 286 (5440), 782–785.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8 (3), 275–282.

Kelley, L.A., MacCallum, R.M., Sternberg, M.J., 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J. Mol. Biol. 299 (2), 499–520.

Lupas, A., Flanagan, J.M., Tamura, T., Baumeister, W., 1997. Self-compartmentalizing proteases. Trends Biochem. Sci. 22 (10), 399–404.

Lupas, A.N., Martin, J., 2002. AAA proteins. Curr. Opin. Struct. Biol. 12 (6), 746–753.

Nobrega, F.G., Nobrega, M.P., Tzagoloff, A., 1992. BCS1, a novel gene required for the expression of functional Rieske iron–sulfur protein in *Saccharomyces cerevisiae*. EMBO J. 11 (11), 3821–3829.

Schultz, J., Milpetz, F., Bork, P., Ponting, C.P., 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc. Natl. Acad. Sci. USA 95 (11), 5857–5864.

Swaffield, J.C., Purugganan, M.D., 1997. The evolution of the conserved ATPase domain (CAD): reconstructing the history of an ancient protein module. J. Mol. Evol. 45 (5), 549–563.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22), 4673–4680.

Tomoyasu, T., Yuki, T., Morimura, S., Mori, H., Yamanaka, K., Niki, H., Hiraga, S., Ogura, T., 1993. The *Escherichia coli* FtsH protein is a prokaryotic member of a protein family of putative ATPases involved in membrane functions, cell cycle control, and gene expression. J. Bacteriol. 175 (5), 1344–1351.

Van de Peer, Y., Frickey, T., Taylor, J., Meyer, A., 2002. Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene 295 (2), 205–211.

Volker, C., Lupas, A.N., 2002. Molecular evolution of proteasomes. Curr. Top. Microbiol. Immunol. 268, 1–22.

Wolf, S., Nagy, I., Lupas, A., Pfeifer, G., Cejka, Z., Muller, S.A., Engel, A., De Mot, R., Baumeister, W., 1998. Characterization of ARC, a divergent member of the AAA ATPase family from *Rhodococcus erythropolis*. J. Mol. Biol. 277 (1), 13–25.

Bibliography:

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

Amores A., Force A., Yan Y.-L., Joly L., Amemiya C., Frity A., Ho R.K., Langeland J., Prince V., Wang Y.-L., Westerfield M., Ekker M., Postlewait J.H. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282:1711-1714.

Beyer A. 1997. Abstract sequence analysisof the AAA protein family. *Protein Sci.* 6:2043-2058.

Brochier C., Philippe H. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* 417:244.

Cavalier-Smith T. 2004. Only six kingdoms of life. *Proc. R. Soc. Lond. B. Biol. Sci.* 271:1251-62.

Cohen G.N., Barbe V., Flament D., Galperin M., Heilig R., Lecompte O., Poch O., Prieur D., Querellou J., Ripp R., Thierry J.C., Van der Oost J., Weissenbach J., Zivanovic Y., Forterre P. 2003. An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi. *Mol. Microbiol.* 47:1495-512.

Coles M., Djuranovic S., Soeding J., Koretke K., Frickey T., Truffault V., Martin J., Lupas A.N. What is the Fold of Abrb-Like Transcription Factors? *Structure* (in press)

Dayhoff M.O., Schwatz R.M., Orcutt B.C. 1978 A model of evolutionary change in proteins. *Atlas of protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington), Vol. 5, Suppl. 3, pp.345-358.

De la Cruz F., Davies J. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* 8:128-133.

Doolittle F.W. 1999 Phylogenetic classification and the universal tree. *Science.* 284:2124-2128

Doolittle R.F and Handy J. 1998. Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr. Opin. Genet. Dev.* 8:630-636.

Doolittle W.F. 1999 Phylogenetic classification and the universal tree. *Science.* 284:2124-2128.

Erdmann R., Wiebel F.F., Flessau A., Rytka J., Beyer A., Froehlich K.U., Kunau W.H. 1991. PAS1, a yeast gene required for peroxisome biogenesis, encodes a member of a novel femily of putative ATPases. *Cell.* 64:499-510.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368-376.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution.* 39:783-791.

Frickey T., Lupas A.N. 2003. Phylogenetic analysis of AAA proteins. *J. Struct. Biol.* 146:2-10.

Frickey T., Lupas A.N. 2004. CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics.* 20:3702-3704.

Frickey T., Lupas A.N. 2004. PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res.* 32:5231-5238.

Froehlich K.U. 2001. An AAA family tree. *J.Cell Sci.* 114:1601-1602.

Gonnet GH, Cohen MA, Benner SA. 1992. Exhaustive matching of the entire protein sequence database. *Science.* 256:1443-5.

Gupta R.S. 1998. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria and eukaryotes. Microbiol. *Mol.Biol. Rev.* 62:1435-1491.

Hendy, M. D., and D. Penny. 1982. Branch and bound algorithms to determine

minimal evolutionary trees. *Mathematical Biosciences*. 59: 277-290.

Henikoff S., Henikoff J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*. 89:10915-10919.

International Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature*. 409:860-921.

International Human Genome Sequencing Consortium 2004. Finishing the euchromatic sequence of the human genome. *Nature*. 431:931-45.

Jaillon O., Aury J.M., Brunet F., Petit J.L., Stange-Thomann N., Mauceli E., Bouneau L., Fischer C., Ozouf-Costaz C., Bernot A., Nicaud S., Jaffe D., Fisher S., Lutfalla G., Dossat C., Segurens B., Dasilva C., Salanoubat M., Levy M., Boudet N., Castellano S., Anthouard V., Jubin C., Castelli V., Katinka M., Vacherie B., Biemont C., Skalli Z., Cattolico L., Poulain J., De Berardinis V., Cruaud C., Duprat S., Brottier P., Coutanceau J.P., Gouzy J., Parra G., Lardier G., Chapple C., McKernan K.J., McEwan P., Bosak S., Kellis M., Volff J.N., Guigo R., Zody M.C., Mesirov J., Lindblad-Toh K., Birren B., Nusbaum C., Kahn D., Robinson-Rechavi M., Laudet V., Schachter V., Quetier F., Saurin W., Scarpelli C., Wincker P., Lander E.S., Weissenbach J., Roest Crollius H. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*. 431:946-57.

Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci*. 8:275-282.

Krause A., Stoye J., Vingron M. 2000. The SYSTERS protein sequence cluster set. *Nucleic Acids Res*. 28:270-272.

Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. 2000. Evolution of two-component signal transduction. *Mol. Biol. Evol*. 17:1956-1970.

Koski L.B., Golding G.B. 2001. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol*. 52:540-542.

Lipman D.J., Pearson W.R. 1985. Rapid and sensitive protein similarity searches. *Science*. 227:1435-1441.

Lupas A.N., Martin J. 2002. AAA proteins. *Curr. Op. Struct. Biol*. 12:746-753.

Martin W. 1999. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *BioEssays*. 21:99-104.

Nelson K.E., Clayton R.A., Gill S.R., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Nelson W.C., Ketchum K.A., McDonald L., Utterback T.R., Malek J.A., Linher K.D., Garrett M.M., Stewart A.M., Cotton M.D., Pratt M.S., Phillips C.A., Richardson D., Heidelberg J., Sutton G.G., Fleischmann R.D., Eisen J.A., Fraser C.M., et al. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature*. 399:323–329.

Ochman H., Lawrence J.G., Groisman E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 405:299-304.

Ohno S. 1967. *Sex Chromosomes and Sex-linked Genes*. Berlin: Springer Verlag. 192 pp.

Ohno S. 1970. *Evolution by Gene duplication*. Springer Verlag, New York, NY.

Proikas-Cezanne T., Wadell S., Gaugel A., Frickey T., Lupas A. N., Nordheim A. 2004. WIPI-1alpha (WIPI49), a member of the novel 7-bladed WIPI protein family, is aberrantly expressen in human cancer and is linked to starvation-induced autophagy. *Oncogene*, 23:9314-9325.

Rabus R., Ruepp A., Frickey T., Rattei T., Fartmann B., Stark M., Bauer M., Zibat A., Lombardot T., Becker I., Amann J., Gellner K., Teeling H., Leuschner W.D., Gloeckner F.-O., Lupas A.N., Amann R., Klenk H.-P. 2004. The genome of Desulfotalea psychrophila, a sulfate reducing bacterium from permanently cold Arctic sediments. *Env. Microbiol*. 6:887-902.

Roelofs J., Van Haastert P.J. 2001. Genes lost during evolution. *Nature*. 411:1013-1014.

Ruepp A., Graml W., Santos-Martinez M-L., Koretke K., Volker C., Mewes W.H., Frishman D., Stocker S., Lupas A.N., Baumeister W. 2000. The genome sequence of the

thermoacidophilic scavenger Thermoplasma acidophilum. *Nature*. 407:508-513.

Saccone C., Gissi C., Lanave C., Pesole G. 1995. Molecular classification of living organisms. *J. Mol. Evol.*. 40:273-9.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.

Salzberg S.L., White O., Peterson J., Eisen J.A. 2001 Microbial genes in the human genome: lateral transfer or gene loss? *Science*. 292:1903-1906.

Santos L., Frickey T., Peters J., Baumeister W., Lupas A., Zwickl P. 2004. Thermoplasma acidophilum TAA43 is an archaeal member of the eukaryotic meiotic branch of AAA ATPases. *Biol. Chem*. 385(11):1105-11

Sicheritz-Ponten T., Andersson S.G. 2001. A phylogenomic approach to microbial evolution. *Nucleic Acids Res*. 29:545-52.

Sokal R.R., Michener C.D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28:1409-1438.

Stanhope M.J., Lupas A.N., Italia M.J., Koretke K., Volker C., Brown J.R. 2001. Phylogenetic analyses do not support horizontal gene transfer from bacteria to vertebrates. *Nature*. 411:940-944.

Stephens S.G., 1951. possible significance of duplications in evolution. *Adv. Genet.* 4:247-265.

Strimmer K., von Haeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13:964-969.

Studier J.A., Keppler K.J. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* 5:729-731.

Swaffield J.C., Purugganan M.D. 1997. The evolution of the conserved ATPase domain (CAD): reconstructing the history of an ancient peptide module. *J. Mol. Evol.* 45:549-563.

Taylor J.S., Braasch I., Frickey T., Meyer A., Vand de Peer Y. 2003. Genome duplication, a trait shared by 22 000 species of ray-finned fish. *Genome Res.* 13:382-390.

Van de Peer Y., Frickey, T., Taylor J.S., Meyer A. 2002 Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. *Gene.* 295:205-211.

Woese C.R. 2000. Interpreting the universal phylogenetic tree. *Proc. Natl. Acad. Sci. USA*. 97:8392-8396.

Woese C.R., Kandler O., Wheelis, M.L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA.* 87:4576-4579.

Wolf S., Nagy I., Lupas A., Pfeifer G., Cejka Z., Muller S.A., Engel A., De Mot R., Baumeister W. 1998. Characterization of ARC, a divergent member of the AAA ATPase family from Rhodococcus erythropolis. *J. Mol. Biol.* 277:13-25.

Zuckerkandl E., Pauling L. 1962. Molecular disease, evolution and genetic heterogeneity. *Horizons in biochemistry* (Academic press), pp. 189-225.