Identification of Conserved Regions and Determination of Relationships in Protein Families by Self-Organizing Maps

Isabelle M. Grimmenstein[1], Miguel A. Andrade[2] and Wolfgang Urfer[1]

[1]Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany
[2]European Molecular Biology Laboratory, D-69117 Heidelberg, Germany

Abstract

The protein family of septins is analyzed with the SOM methodology of Andrade et al. (1997) to determine the family relationships and the key residues responsible for the obtained classification. These key residues are candidates for determining functional sites of the proteins. The advantages of the applied SOM methodology compared to other methods are discussed. Its limitations and drawbacks are pointed out as well. Finally, possible enhancements of the methodology for future research are given.

1 Introduction

Proteins are the building blocks of life. On one hand, they form the structural fabric for every organism and, on the other hand, they are the basis for all the biochemical processes in the cell. Therefore, the investigation of proteins is essential to get a better understanding of the processes of life. The analysis of protein family data can especially give a better insight into biological relationships where the family members are supposed to have diverged through mutational events (insertions, deletions or substitutions of amino acids) from a common ancestor.

For drawing conclusions about the functional role of the diverse members of a given protein family it is important to know the family relationships. Further insights can be gained by detecting conserved sites in the proteins of the family. Conserved residues remained the same during the course of evolution or are at least very similar at the corresponding sites in the whole protein family or in special subgroups. These conserved residues, also called key residues, are strong candidates for representing sites with important biological functions responsible for the characteristics of the whole protein family or rather a special subgroup. These functionally important sites can be either catalytic sites where interactions with other molecules take place or they can be determinant for the given three-dimensional structure.

One possibility to obtain information about the inherent relationships in a protein family is the use of phylogenetic methods which estimate a phylogenetic tree from the sequence data like parsimony (Fitch, 1971; Sankoff and Cedergren, 1983), distance methods, e.g. FITCH (Fitch and Margoliash, 1967), neighbor joining (Saitou and Nei, 1987) and UPGMA (*u*nweighted *p*air *g*roup *m*ethod using *a*rithmetic averages, Sokal and Michener, 1958), or *m*aximum *l*ikelihood (ML) methods. All these methods are established, but have also disadvantages. The maximum parsimony approach was shown to be inconsistent (Felsenstein, 1988) and does not correspond to evolutionary mechanisms by trying to minimize the number of substitutions (Goldman, 1996). Distance methods reduce the available information by using pairwise distances between the sequences as input data instead of the original protein sequences and the models for calculating pairwise distances between amino acid sequences are not so well developed yet compared to models for DNA sequences (Goldman, 1996). Neighbor joining and UPGMA have moreover the disadvantage of not giving explicitly an objective optimization function which allows different trees to be compared. Further, both can produce trees containing negative branch lengths which is biologically not meaningful (UPGMA however only, if some of the employed distances are defined as negative). UPGMA assumes additionaly a molecular clock which is in general not satisfied and was shown to give unrealistic results (Huelsenbeck and Hillis, 1993). The ML approach has the advantage of using a well-defined probabilistic model for sequence evolution and optimizes a likelihood function which gives an objective criterion for the assessment of different results. In contrast the other methods are rather heuristically motivated (Felsenstein, 1988). The ML approach should therefore always be preferred when possible(Goldman, 1996). However, the great disadvantage of the ML approach is that it is especially for protein sequences computationally very demanding and hence only applicable to smaller sets of sequences. With the quartet puzzling approach of Strimmer and von Haeseler (1996), which is a heuristic search strategy by forming quartet trees, the maximum likelihood tree can be determined for a larger set of sequences, but it has also its limitations. The results by quartet puzzling are a bit worse than those of the conventional ML search algorithm. In addition, when the investigated sequences are short there might be overfitting because of the many parameters to be estimated in the ML approach. Furthermore, it is not always necessary to estimate a full phylogenetic tree when instead a classification on only special resolution levels is sufficient.

In this paper we use a different approach for the analysis of protein family data, the *S*elf-*O*rganizing *M*aps (SOMs) of Kohonen (1982), modified and applied in the context of protein data by Andrade et al. (1997). With this approach we obtain classifications for proteins of the septin family with different resolutions, construct a tree displaying the family relationships and determine moreover the key residues responsible for the classification.

2 The protein family of septins

The septins are a family of conserved proteins belonging to the superclass of P-loop GTPases (Leipe et al., 2002). This superclass can be divided into two large classes comprising together all protein families which bind and hydrolyze GTP (*g*uanine *t*ri*p*hosphate). They share apart from two other motifs as common motif the P-loop, where the phosphate of the GTP molecule binds. There are altogether over 20 distinct families belonging to the GTPase superclass which can be further subdivided into 57 subfamilies. The septins form one of those subfamilies belonging to the first class which comprises the majority of the well-known GTPases involved in translation, signal transduction, cell motility and intracellular transport.

The proteins of the septin family are represented in varying numbers in a large scale of eukaryotic organisms like fungi, worms, fruit flies, mice and humans, but they are missing in plants (Longtine et al., 1996). In higher organisms they are widespread over different types of tissue indicating an important role in the cell for the septins. In humans for example, they were found in tissues like skin, brain, kidney, muscle, bone marrow, ovarian, uterus and testis. Septins were first discovered in yeast (*Saccharomyces cerevisiae*) in relation with bud growth and cytokinesis about 30 years ago (Hartwell, 1971) and are named for their involvement in forming the septum between two dividing cells, called septation (Field and Kellogg, 1999). Septins were also shown to be involved in cytokinesis in higher eukaryotes indicating a conserved function over different types of organisms. Further studies revealed that some play also a role in other biological processes during the different stages of the cell cycle like vesicle trafficking and vesicular fusion with the cell membrane (Kartmann and Roth, 2001). In yeast, septin proteins are further involved in spore formation. It was also assumed that they could constitute a novel cytoskeletal system because of their ability to form filaments in vitro or that they serve as scaffolds for other proteins in signaling pathways because of their interaction with a wide variety of different proteins (Field and Kellogg, 1999). Kartmann and Roth (2001) discuss beyond a potential role of septins in oncogenesis. One remarkable property of the sequences of this family is that they tend to form complexes between each other. This was shown, for example, for the septins Pnut, Sep1 and Sep2 in *Drosophila melanogaster* (Field et al., 1996), Cdc3p, Cdc10p, Cdc11p and Cdc12p in *Saccharomyces cerevisiae* (Frazier et al., 1998) and CDCrel-1 and KIAA0202 in humans (Blaser et al., 2002). This could explain the expansion of the family. Altogether, there is still uncertainty about their functional roles in the cell and only little is known about their biochemical mechanisms.

The primary structure, i.e. the sequence of amino acids in the protein, is well conserved throughout the family of septins. In general they display a sequence identity of at least 26 % over their entire length and consist of 275 to 539 amino acids (Field and Kellogg, 1999). In the central region, the GTPase domain, the sequence similarity is highest. It contains the three conserved sequence motifs characteristic for P-loop GTPases shown in Figure 1. The first motif constitutes the P-loop (Saraste et al., 1990). The N- and C-terminal regions

	GxxxxGK[S/T] DTPG xKxD	
N-terminus	*GTPase domain*	*C-terminus (coiled-coil)*

Figure 1: Sequence structure of septins with three conserved motifs (cf. Field and Kellogg, 1999; Momany et al., 2000). The middle motif is already more specified for the septins, which is in the general form given by DxxG. (The coding for the amino acids is listed in Table 1.)

vary considerably in length and sequence composition through the different members of the septin family and are even missing in some. The C-terminus is predicted to form a coiled-coil domain, a structure of two helices fitted into each other, thought to be involved in protein-protein interactions (Longtine et al., 1996). The overall three-dimensional structure of the septins is so far unknown.

The analysis with the self-organizing network methodology should bring more insights into the septins by clarifying their internal organization and detecting key residues, which are candidates for functional important regions in the proteins. This knowledge could help for future research in the design of genetic experiments, in the determination of the three-dimensional structure and in the investigation of diseases.

3 The SOM algorithm

For the application of the SOM methodology to protein data, like in other phylogenetic methods, the N considered amino acid sequences must be given in a multiple alignment where the amino acids of the sequences are arranged by introducing gaps in such a way that amino acids with the same evolutionary origin should have the same position. The gaps are accounting for insertions and deletions of amino acids occurred through mutational events during the course of evolution. All sequences in the alignment have the same length L.

To obtain the aligned proteins in a numerical treatable form the sequences are binary coded. Every position of the aligned sequences is described by 20 components according to the 20 amino acids found in proteins. A "1" is assigned to that component which corresponds to the amino acid in the considered position and a "0" is assigned to the remaining 19 components (see Figure 2 and Table 1). In case of a gap all the 20 components are assigned a "0". This results in sequence vectors x_n, $n = 1, \dots, N$, of dimension $20L$.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2: Binary coding of a sequence position with amino acid D (Aspartic Acid).

A	Alanine	M	Methionine
C	Cysteine	N	Asparagine
D	Aspartic acid	P	Proline
E	Glutamic acid	Q	Glutamine
F	Phenylalanine	R	Arginine
G	Glycine	S	Serine
H	Histidine	T	Threonine
I	Isoleucine	V	Valine
K	Lysine	W	Tryptophan
L	Leucine	Y	Tyrosine

Table 1: The one-letter code for the 20 amino acids.

The high-dimensional sequence vectors $x_n \in \mathbb{R}^{20L}$, $n = 1, \ldots, N$, are projected by the iterative SOM algorithm (Kohonen, 1982 and 2001; Andrade et al., 1997) onto a two-dimensional map giving a clustering and a visualization of the inherent structure of the sequence vectors. Neighborhoods in the map reflect thereby usually similarities in the sequence vectors.

The map is a rectangular lattice $\mathcal{L}$ (cf. Fig. 3) of predetermined size $a \times b$ with m vertices P_i, $i = 1, \ldots, m$, where $m = ab$, i.e. the map is given by the set

$$\mathcal{L} = \{P_i = (i_1, i_2) | i_1 = 1, \ldots, a; i_2 = 1, \ldots, b\} \quad \subset \quad \mathbb{R}^2. \tag{1}$$

A weight vector w_i for $i = 1, \ldots, m$ is assigned to each vertex P_i of the map having the same dimensionality $20L$ as the sequence vectors. In contrast to the sequence vectors the components of the weight vectors can take any real value between 0 and 1. The components of the weight vectors are set initially to random values and then adapted during the training algorithm of the SOM procedure to the given data. At the end of the procedure different weight vectors should represent different subsets of the sequence vectors. The training procedure of the SOM is given as follows with t as discrete time variable:

Training:

Initialization ($t = 0$): a) Set the components w_{ij} of the weight vectors w_i to arbitrary values between 0 and 1, i.e. $0 \leq w_{ij} \leq 1$, $i = 1, \ldots, m$ and $j = 1, \ldots, 20L$.
b) Choose as current data set $D(t)$ the complete set of sequence vectors $\mathcal{D} = \{x_1, \ldots, x_N\}$.

Updating of the weight vectors w_i ($t > 0$): Select an arbitrary sequence vector x_n from the current data set $D(t)$. Update all the weight vectors w_i for the selected sequence vector x_n according to

$$w_i(t+1) = w_i(t) + \alpha K_{i^*i}(t) \left(x_n - w_i(t)\right), \qquad i = 1, \ldots, m, \tag{2}$$

with

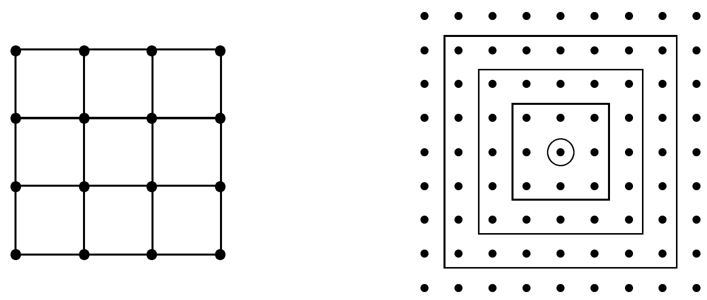

Figure 3: Display of the vertices in an example lattice of size 4×4 and neighborhood areas of size 3×3, 5×5 and 7×7. In the second picture, the grid lines are omitted for a clear display and the central vertex of the neighborhood areas is highlighted by a circle (cf. Kohonen, 2001).

$i^* = \arg\min_i \|x_n - w_i(t)\|$ (euclidean distance),
α: constant learning parameter ($0 < \alpha < 1$),
$K_{i^*i}(t)$: neighborhood function with $K_{i^*i}(t) \to 0$ when $t \to \infty$, defined by

$$K_{i^*i}(t) := K\left(\delta(P_{i^*}, P_i); t\right) := K_t(d) := \begin{cases} 1, & d = 0 \\ f_t(d), & 0 < d \leq r_t \\ 0, & d > r_t \end{cases}$$

where

$\delta(P_{i^*}, P_i) =: d$ defines the distance between the vertices P_{i^*} and P_i,
$f_t(d)$ is a monotonously decreasing function in d, and
r_t defines the neighborhood size, decreasing with time t ($\lim_{t\to\infty} r_t = 0$).

Remove sequence vector x_n from $D(t)$ and set as new data set $D(t) \setminus \{x_n\}$ or the complete set of sequence vectors $\mathcal{D}$ if no sequence is left in the set (onset of a new training cycle).

Termination: The procedure is terminated when convergence is assumed for the weight vectors w_i and the neighborhood area contains only the central vertex.
The sequence vectors x_n, $n = 1, \ldots, N$, are assigned to that vertex P_{i^*} on the map with the closest weight vector w_{i^*}, i.e.

$$\|x_n - w_{i^*}(t)\| = \min_i \|x_n - w_i(t)\|.$$

The updating procedure makes the weight vectors move closer to the presented sequence vector x_n in a training run according to the selected learning parameter α. The size of the parameter α influences the speed of the training procedure. In usual SOM implementations, α is chosen as a time decreasing factor, but Andrade et al. (1997) suggest to keep α constant during the whole training procedure to avoid convergence of the weight vectors caused by the learning parameter and not by the data themselves.

Not all the weight vectors are updated to the same extent in a training run. There is only one weight vector to which the full α factor in the updating step is applied. It is the weight vector w_{i^*} having the smallest distance to the presented sequence vector x_n, here measured as euclidean distance. For a topographic ordering of the sequence data the weight vectors of the neighboring vertices are as well updated, but with decreasing extent by increasing distance from the central vertex holding the closest weight vector. Weight vectors from vertices lying outside the neighborhood area are not updated at all. The extent to which the weight vectors are updated in a training run is controlled by the monotonously decreasing neighborhood function $K_t(d)$. We choose, in analogy to Andrade et al. (1997), the neighborhood function

$$K_t(d) = \exp\left\{\ln(0.1)\, d/r_t\right\}, \tag{3}$$

where r_t defines the size of the neighborhood area in dependence of the time t by giving the horizontal distance between the central vertex to the border of the neighborhood and d indicates the distance between the central vertex and the vertex holding the weight vector to be updated. The distance between the two vertices P_{i^*} and P_i is determined by the euclidean distance measure, i.e.

$$d = \delta(P_{i^*}, P_i) = \sqrt{(i_1^* - i_1)^2 + (i_2^* - i_2)^2}. \tag{4}$$

To achieve convergence for the weight vectors w_i, the size r_t of the neighborhood is gradually shrunk through the training cycles. A training cycle is thereby completed when every sequence vector is presented once to the system in random order. It comprises therefore N training runs. The random presentation of the sequence vectors is necessary to add noise to the system for omitting subobtimal classifications. Here, we use square shaped regions of varying size as neighborhood areas (cf. Fig. 3). The SOM procedure is started with a neighborhood region covering the whole map and ends with a neighborhood area consisting of only the central vertex. In usual SOM implementations, a constant rate of decrease is chosen by trial and error for the neighborhood size. Andrade et al. (1997) suggest in contrast a different procedure, followed also here, which makes the shrinkage tailored to the given data.

After each training cycle c a dispersion value $s_i(c)$ is computed for every weight vector w_i over the last γ training cycles:

$$s_i(c) = \sqrt{\sum_{t=(c-\gamma)N}^{cN} \frac{|\bar{w}_i(c) - w_i(t)|^2}{\gamma}}, \tag{5}$$

with $\bar{w}_i(c) = \frac{1}{\gamma N}\sum_{t=(c-\gamma)N}^{cN} w_i(t)$ as the mean of the w_i values at vertex i over the last γ training cycles.

To measure the dispersion over the whole map regarding the last γ training cycles the value

$$\bar{s}(c) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} s_i(c)} \tag{6}$$

is calculated. Convergence is assumed for the weight vectors and hence the neighborhood size decreased, when the condition

$$\frac{|\bar{s}(c) - \bar{s}(c-1)|}{\bar{s}(c)} < \sigma \tag{7}$$

is fulfilled for a given threshold σ, i.e. when the degree of change of the weight vectors over the last γ training cycles is sufficient small.

Proteins whose sequence vectors are assigned at the end of the SOM procedure to the same vertex form a cluster. However, not all vertices are assigned sequence vectors. The weight vectors give a summary of the corresponding classes and their components hold information about potential key residues.

4 Tree construction

The application of the SOM algorithm with a predefined map size $a \times b = m$ can yield only one special classification of the considered amino acid sequences in not more than m clusters. This is not able to reflect the whole information inherent in the protein data. However, one can combine the clustering results received with maps of different resolution level for constructing a tree which displays the family relationships within the protein data. This is done by linking clusters which divide on a higher resolution level into subclusters.

Ideally, the separation of a cluster into subclusters should be unambiguous. However, sometimes it happens that two sequences clustered at one level in two separate clusters are assigned on the next higher resolution level to a joint cluster. Such a collapse leads to an inconsistency and must be resolved afterwards. There are two possibilities to resolve a collapse which occurs from level i to level $i+1$:

A. Classify the affected sequences also at the preceding level i in one joint cluster or

B. assign the affected sequences also at the following level $i+1$ to two separate clusters.

To decide between the two possibilities the following procedure is carried out for two collapsing sequences A and B (cf. Andrade et al., 1997):

1. Determine at level $i+1$ which of the two sequences is mapped worse to the common cluster, i.e. which one has the larger distance to the corresponding weight vector.
   (Let this be sequence A without loss of generality.)

2. Compute the mapping score of sequence A at level i and level $i+1$.
   The mapping score of a sequence is the rank of its distance to its corresponding weight vector in comparison to the distances of all the other sequence vectors (built in a decreasing order so that the best mapped sequence gets the score 1).

3. Keep the clustering on that level where sequence A has the lower, i.e. better mapping score.
   If the mapping score of A is worse on level i, sequence A is joined to the cluster containing sequence B (solution A).
   If the mapping score of A is worse on level $i+1$, split sequence A from the joint cluster by placing it in a newly generated one (solution B).

The occurring collapses are resolved in several passes over the tree, beginning at the lowest level towards the tip of the tree. In the case of more than two sequences involved in a collapse, the sum of the mapping scores is considered for decision.

To measure the reliability of a tree Andrade et al. (1997) suggest as a criterion the collapse index, defined by

$$\frac{\#collapses}{\#sequences \cdot \#levels}. \tag{8}$$

Some sequences might be difficult to classify at all which should be indicated by a high number of collapses for the respective sequences.

The described methodology gives a rooted tree which differs in some aspects from regular phylogenetic trees. First the branch lengths do not represent evolutionary distances like in phylogenetic trees. Furthermore, the leaves are not necessarily consisting of only one sequence. The number of sequences at the leaves depends on the selected resolution levels for the employed maps and on the family relationships. Even with many resolution levels closely related sequences may stay together in one cluster.

5 Results

We applied the presented SOM methodology to proteins of the septin family to derive family relationships and to determine key residues. For some proteins of the septin family, Momany et al. (2000) considered already the family relationships by constructing a tree with the neighbor joining algorithm (Saitou and Nei, 1987). They took however only fungal septins into account. In contrast, we are interested in the family relationships of septins belonging to different species. Leipe et al. (2002) explored also the family relationships of septins but only in the context of the large GTPase family and not in detail.

For our analysis, we used a multiple alignment of 98 septins from different species like humans, rats, mice, fruit flies and yeasts, derived by the SP-trEMBL data base (http://igbmc.u-strasbg.fr:8080/DbClustal/dbclustal.html) with a fragment of the innocent bystander protein (SEP1_DROME) from the fruit fly (*Drosophila melanogaster*) as query sequence. The alignment consisted of 1019 positions. For the construction of a tree, maps with 6 different resolution levels were combined (a 2×1, 3×1, 2×2, 3×3, 4×4 and a 5×5 map). The parameters which had to be determined in advance were set, analogous to the experiences of Andrade et al. (1997), to the following values:

- $\alpha = 0.1$ for the learning parameter,
- $\gamma = 5$ for the number of cycles taken into account for computing the dispersion value and
- $\sigma = 0.005$ as threshold for defining convergence of the weight vectors.

We found in our trees that the considered sequences of the septin family separate into four clear groups. However, by running the algorithm with different initializations, we received always trees which differed in some aspects from the other ones. The four groups stayed apart from some deviations essentially the same, but the way of splitting up of the clusters and the levels where the splittings occurred differed. Two example trees are given in Figure 4 and 5. The four groups in the two trees are different numbers of sequences assigned. In the tree of Figure 4 they contain top down 41, 23, 13 and 21 sequences. The corresponding groups in the tree of Figure 5 contain similar numbers of sequences, namely 38, 20, 19 and 21 sequences. On the highest resolution level there are 26 clusters in the tree of Figure 4 and with 27 clusters just one more in the tree of Figure 5.

We give a brief interpretation for the four groups derived from the tree of Figure 5, which we retrieved by a check in the SWISS-PROT protein sequence data base (Bairoch and Apweiler, 2000) with its cross references. The first group (going from node 3) contains only septins from humans, mice and rats being found in different tissues like skin, brain, ovarian and breast. The function of the septins in this group is mainly unclear. The second group (going from node 7) comprises all the fungal septins in the data set, like from *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Candida albicans* and *Emericella nidulans*, the major part of which is involved in cytokinesis. The group further contains some septins from higher organisms, namely 3 septins from mice, 2 from humans, 1 from rats, 1 from fruit flies (*Drosophila melanogaster*) and 1 from *Caenorhabditis elegans* (a nematode) also partly involved in cytokinesis. The third group (going from node 8) covers mainly septins from humans, mice and fruit flies. Moreover, there were 2 septins from rats, and 1 septin from each of the following: *Macaca fascicularis* (a monkey), *Xenopus laevis* (African clawed frog) and *Geodia cydonium* (a sponge). Most of these septins are involved in cytokinesis and represented in the brain (apart from other tissues). The fourth group (going from node 2) contains only septins from mice, humans and fruit flies from different types of tissues like heart, brain, muscle, embryo, uterus and bone marrow. Their functions are unclear. Some are thought to be involved in cytokinesis.

The number of collapses differed considerably from tree to tree and hence the collapse index, because we considered only trees for the same set of sequences

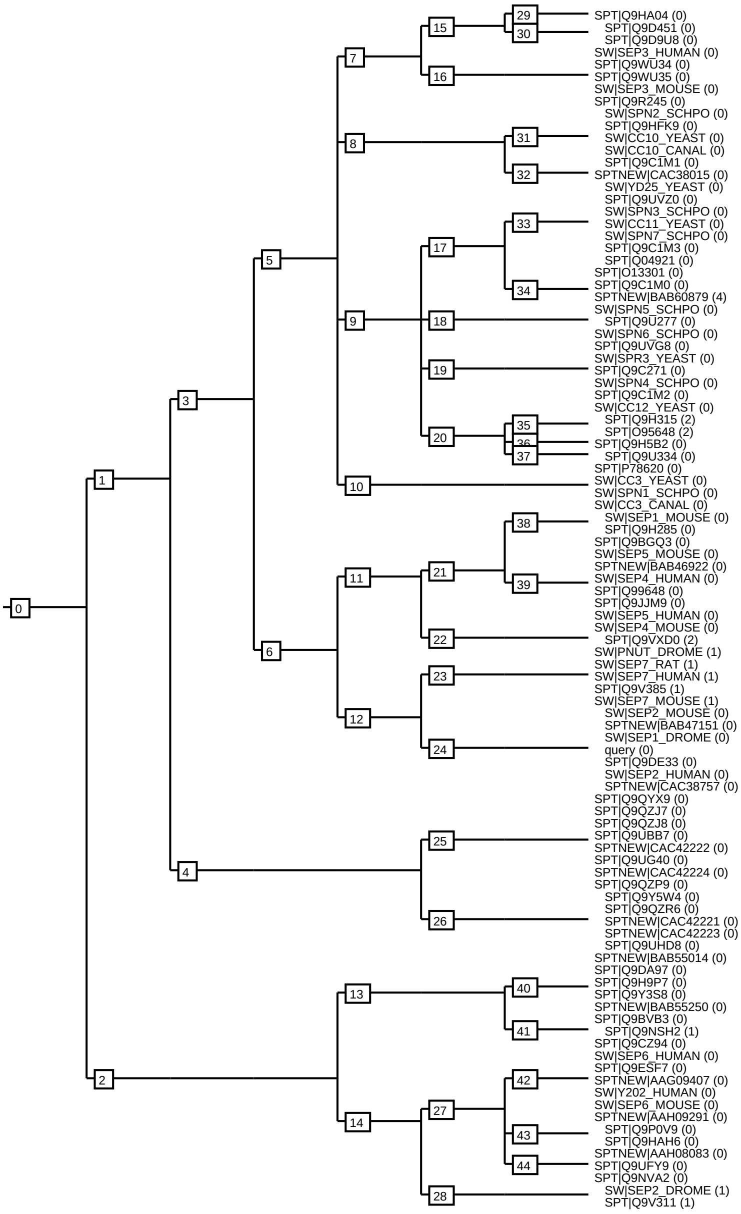

Figure 4: Classification tree derived for 98 septins by the SOM approach. The number of collapses is given in parenthesis behind each sequence name. There are altogether 18 collapses. The nodes are numbered from left to right. The four detected groups go off from node 2, 4, 5 and 6.

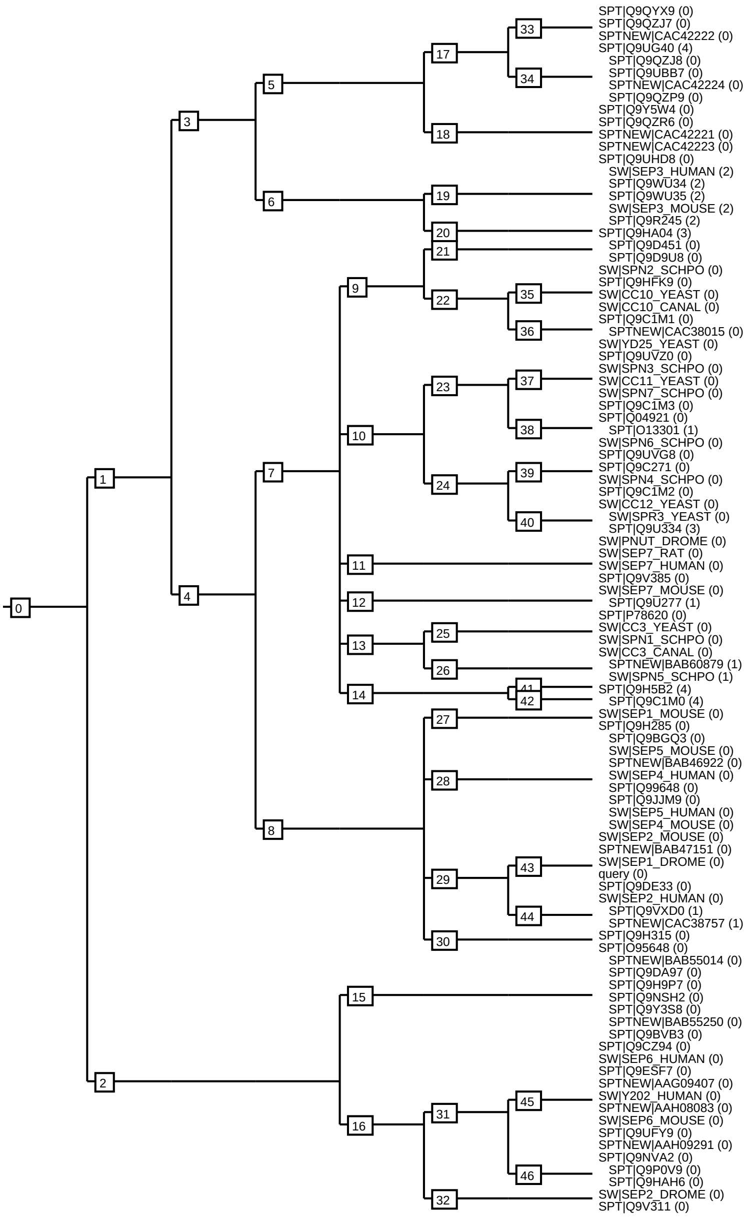

Figure 5: Alternative classification tree derived for 98 septins by the SOM approach. The number of collapses is given in parenthesis behind each sequence name. There are altogether 34 collapses. The nodes are numbered from left to right. The four detected groups go off from node 2, 3, 7 and 8.

with the same number of levels. We experienced further that a tree with a lower collapse index is not giving necessarily a better classification for the analyzed sequences. For example, the two sequences SPTNEW|CAC42222 and SPT|Q9HA04 should be classified near to each other in the trees, because they are closely related. The sequence SPT|Q9HA04 is a fragment (with a missing C-terminus) of SPTNEW|CAC42222 with an overall sequence identity of 56.8 % between the two. However, in the tree of Figure 4 with just 18 collapses altogether, these two sequences are assigned to two different clusters (going from node 4 and 5), whereas in the tree of Figure 5 the two sequences are classified properly in the same cluster (going from node 3) though it has clearly more collapses, namely 34 altogether. The inherent family relationships in the given sequences are therefore not always represented correctly in a tree derived by the SOM procedure and the collapse index seems not to be a sufficient criterion for the assessment of an obtained classification.

However, proteins with an identical amino acid sequence were always classified correctly together in the same cluster in the considered cases. In our data set we had four pairs of identical sequences, namely

SW|Y202_HUMAN	and	SPTNEW|AAG09407,
SPT|Q9NVA2	and	SPTNEW|AAH08083,
SW|SEP4_HUMAN	and	SPTNEW|BAB46922,
SPT|Q9Y5W4	and	SPTNEW|CAC42223.

The sequences themselves which had the collapses differed over the diverse trees we considered. There were some sequences which had collapses in several trees, but with varying numbers and none of them had collapses in all the trees we compared more thoroughly with each other (altogether 8 trees). The protein SPT|Q9UG40 is the sequence in the data set which had the most collapses over the compared trees. It had collapses in 7 out of the 8 trees, always with the highest occurring numbers of 3 or 4 collapses. The two proteins SPT|Q9HA04 and SPT|Q9U334 had the second frequent number of collapses over the trees. Both had in 6 out of the 8 trees collapses with varying numbers from 1 to 4. The high number of collapses can indicate a special role for these three sequences in the septin family, making them difficult to classify unambiguously. An enquiry in the SwissProt date base revealed that the protein SPT|Q9U334 is not a septin at all and was wrongly joined to the data set. It belongs to the peptidase family. The two other proteins are hypothetical proteins with unknown function so far, whereas SPT|Q9UG40 constitutes, like SPT|Q9HA04, another fragment of the protein SPTNEW|CAC42222. In the two example trees given in Figure 4 and 5, the sequences SPTNEW|BAB60879 and SPT|Q9VXD0 are the only ones which have respectively one collapse in each tree. All the other 10 resp. 14 sequences have just collapses in the first resp. second tree. Overall, it seems difficult to distinguish between sequences having collapses just by chance or because of a real ambiguity without comparison of several trees. In general, it is more difficult to assess the relationships of the sequences to each other in a tree generated by the SOM approach compared to phylogenetic trees, because the branches do not display evolutionary distances.

By analyzing the obtained clusters from the tree for each node and alignment position we determined conserved residues of the sequences. We distinguished between residues which are conserved for a special node being therefore tree-determinant and residues which have been already conserved in the parent node. An outline of the determinant residues around the P-loop is displayed in Figure 6 for the tree of Figure 5.

Positions with gaps in a cluster of an associated node were treated as jokers in the determination of the conserved residues. This means that an amino acid is also displayed as conserved when only one sequence of the cluster contains at the considered position that amino acid and all the other sequences have a gap there. Additional information can however be gained in this regard by comparing the conserved residues from other nodes. The alternative of not taking into account any position containing gaps would not be appropriate in the given case, because no alignment position would be left for consideration due to the use of whole amino acid sequences and the inclusion of fragments. An intermediate solution, where positions are taken as gaps when a predefined, proportion of sequences having a gap in the considered positions is exceeded, is also not reasonable, because of the cumulative character of the clusters in the tree. The proportion of gaps might change from level to level and therefore also the assignment of a gap to a node and its daughter nodes.

In the following we give an overview of the main results regarding the conserved residues related to the tree of Figure 5. We found several positions over the entire alignment of 1019 positions which are quite well conserved in the whole family, i.e. were represented by the same amino acid for almost every node, especially in the central part. These are for example in position 292 amino acid V, in 300 a Q, in 309 a G, in 332 a T, 337 an L, in 238 an F, in 434 an L, in 471 an E, in 476 an R, in 485 an R, in 522 a G, in 544 a G, in 546 a T and in 669 an R. The position 647 with a P and the position 673 with a W are conserved the best in the tree, having only two nodes not classified as conserved because of gaps. The reason why not all nodes are displayed as conserved is, above all, because the data set contains several protein fragments which are missing partly the central motives. The detected conserved residues might be important for the general function of the family, but they are not the tree-determinants of the family. The three motives typical for P-loop GTPases and introduced in chapter 2 were also retrieved as conserved over the whole set of sequences. The P-loop motif GxxxxGK[S/T] turned out to be more specific for the considered sequences, given by G[E/Q][T/S]GLGKS. The amino acid Q in the second position of the P-loop is thereby characterizing the protein group at node 3 whereas the amino acid E is characteristic for the two groups at node 2 and 8. In the third position of the P-loop, the amino acid T is characteristic for the node 2 opposed to the nodes 3 and 8 for which amino acid S is characteristic. The DxxG motif, already determined more specifically as DTPG for septins, was approved by our findings. The xKxD motif could be determined more precisely as AKAD.

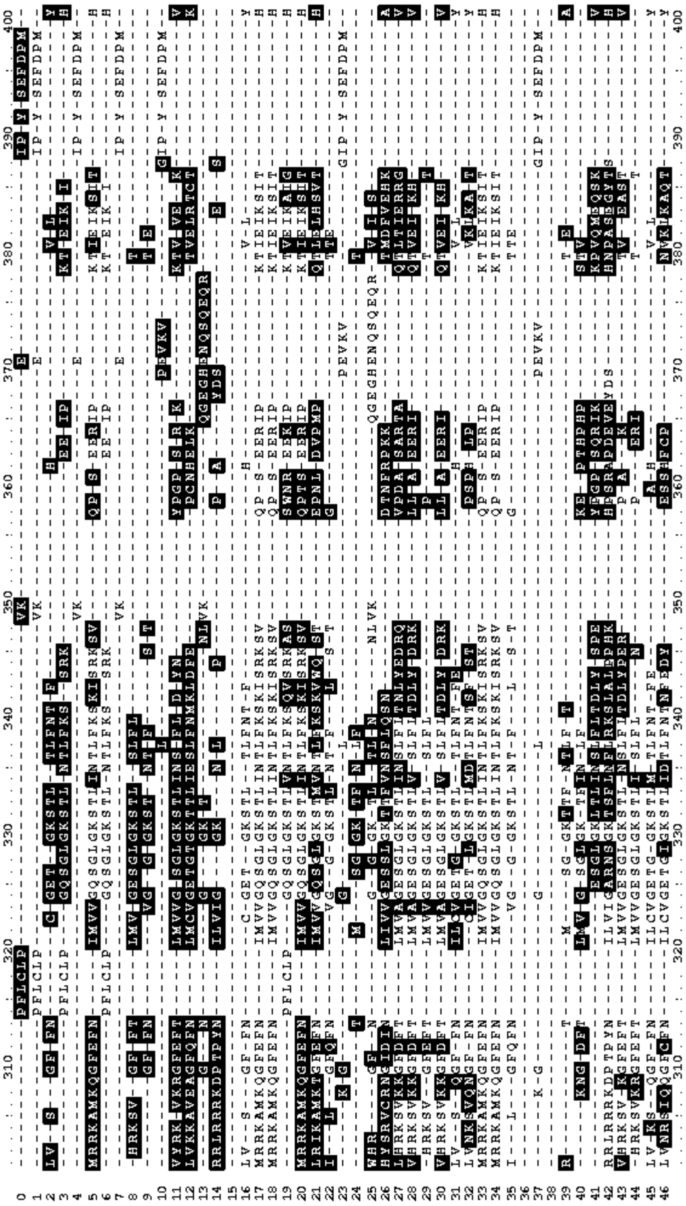

Figure 6: Outline of the conserved residues determined for each node of the tree displayed in Figure 5 from alignment position 301 to 400. It contains the P-loop from position 324 to 331. The horizontal axis indicates the alignment position, the vertical axis the node number in the tree. Amino acids with black background are conserved in the associated node, i.e. are tree-determinant. Amino acids with no background are already conserved in the parent node.

We found also positions in the alignment which are conserved especially in the four derived groups but not in the whole family and are therefore tree-determinant. In position 339 we found for node 2 amino acid N as conserved, for node 3 amino acid K and for node 8 amino acid L. In position 437 we found for node 2 an L, for node 3 a V and for node 8 an S. In position 641 we found for node 2 an M, for node 3 an R, for node 7 at higher levels an A respectively an L and for node 8 a K. Also the second and third position of the P-loop motif are determinant for the tree as already presented above.

6 Limitations of the SOM methodology and possible enhancements

We applied the modified SOM approach by Andrade et al. (1997) to proteins of the septin family and received a classification into four main groups. All the fungal septins in the data set were thereby classified together into the same group and all the other three groups comprised only septins from higher organisms like humans, mice and fruit flies. This is biologically meaningful from the aspect that the members of those two eukaryotic groups belong to two different kingdoms, namely the fungi and the animals (metazoa), having separated from each other already in the early stages of evolution.

We examined further sequences which were difficult to classify and detected a protein which was wrongly assigned to the data set and not belonging to the family of septins at all. However, this was only reliably possible with the comparison of the results from different runs of the algorithm.

We determined in addition conserved residues for the whole data set and residues which were characteristic for the four derived groups. We retrieved among others the three motives typical for P-loop GTPases as conserved and were able to specify them for the considered set of septins.

As already indicated in the previous section, we found several drawbacks of the SOM methodolgy applied to our protein data and we want to list them briefly.

Limitations:

- Different runs of the SOM procedure with different initializations yield different results.
- The collapse index is not enough for the assessment of a tree. It does not tell anything about the quality of the found clustering itself.
- There is no global optimization criterion for the assessment of a certain mapping with a predefined resolution level and correspondingly none for a whole tree.
- It is not assured that the weight vectors will converge.
- The neighborhood preservation is not guaranteed.

- The tendency of the SOM algorithm as pointed out by Andrade et al. (1997) to produce clusters with equal size seems in the context of phylogenetically related proteins not justified.
- There is no evolutionary model used.
- The euclidean distance does not account for the biochemical properties of the protein sequences. By using the euclidean distance only identical residues are taken into account and mutations to residues with similar properties are discarded.
- The accuracy of the phylogenetic estimates depends strongly on the quality of the multiple alignment. Bad alignments will lead to bad results with the SOM procedure.
- The display of the family relationships in tree form is not appropriate in the case of horizontal transfer of genetic material.

We want to note that the whole SOM procedure, as employed here with a decreasing neighborhood area over the training cycles, is not optimizing a quality criterion. However, Bock (1998) gives a good overview that the SOM procedure with a non time-decreasing neighborhood function minimizes a finite sample clustering criterion as a generalized stochastic approximation approach. Thus, the applied procedure minimizes different finite sample criteria in sections.

To overcome the limitations of the SOM methodology in the context of the analysis of protein family data, we propose some enhancements which will be the topic of future research.

Enhancements:

- For the assessment of a tree in comparison with others one should introduce apart from the collapse index further criteria which measure the quality of the clusters themselves, i.e. how homogeneous the individual clusters are and how heterogeneous among each other.
- Additionally, we propose to use the GTM (*g*eneral *t*opographic *m*apping) approach of Bishop et al. (1998) instead of the SOMs by Kohonen. This algorithm is based on modelling the distribution of the data in terms of latent variables. Such latent variables are used by Márkus et al. (1999) to analyze spatially and temporally correlated data. The GTM gives also a topographic mapping of the given high-dimensional data onto a low-dimensional space, but in contrast it has a probabilistic background and maximizes a likelihood function as global optimization criterion.
- To reduce the dependency of the quality of the results on the quality of the given alignment an iterative procedure between the tree construction and the improvement of the alignment could be designed.

Finally we conclude that the SOM or GTM approach is not suited for constructing a complete phylogenetic tree. For this purpose, standard phylogenetic methods should be preferred like maximum likelihood methods, FITCH or neighbor joining. However, they are tools to obtain biologically meaningful classifications for a given protein family and to detect key residues potentially responsible for characteristics of the found subgroups or the whole family. To derive subgroups it is not necessary to compute a full blown-up phylogenetic tree. Especially in the case when the multiple alignment consists of many short sequences, a phylogenetic tree might lead to overfitting.

Other possible areas for application of the SOM or GTM methodology are in biotechnology, such as cDNA microarrays and high-density oligonucleotide chips. These tools allow simultaneous monitoring of the expression of thousands of genes under different conditions. Guimarães and Urfer (2002) recommend the use of regression-type models of SOMs (Bock, 2000) for the analysis of such gene expression data.

7 Acknowledgements

We would like to thank Dr. Ingrid Vetter and Prof. Alfred Wittinghofer from the Max-Planck-Institute for Molecular Physiology in Dortmund for their support concerning the biological background. Also we wish to thank Dr. Dirk Husmeier from Biomathematics and Statistics Scotland (BioSS) in Dundee for the helpful discussions and comments.
This work has been supported by the German Research Council (DFG) through the Graduate College and the Collaborative Research Center "Reduction of Complexity for Multivariate Data Structures" (SFB 475) at the University of Dortmund.

References

ANDRADE, M.A.; CASARI, G.; SANDER, C. and VALENCIA, A. (1997): Classification of Protein Families and Detection of the Determinant Residues with an Improved Self-Organizing Map. *Biological Cybernetics, 76, 441–450.*

BAIROCH, A. and APWEILER, R. (2000): The SWISS-PROT Protein Sequence Database and its Supplement TrEMBL in 2000. *Nucleic Acids Research, 28, 45–48.*

BISHOP C.M.; SVENSEN, M. and WILLIAMS, C.K.I. (1998), GTM: The Generative Topographic Mapping, *Neural Computation, 10, 215–234.*

BLASER, S.; JERSCH, K.; HAINMANN, I.; WUNDERLE, D.; ZGAGA-GRIESZ, A.; BUSSE, A. and ZIEGER, B. (2002): Human Septin-Septin Ineraction: CDCrel-1 Partners with KIAA0202. *FEBS Letters, 519, 169–172.*

BOCK, H.H. (1998): Clustering and Neural Networks. In: A. Rizzi, M. Vichi and H.H. Bock (Eds.): *Advances in Data Science and Classification.* Springer, Berlin, 265–277.

BOCK, H.H. (2000): Regression-Type Models for Kohonen's Self-Organizing Networks. In: R. Decker and W. Gaul (Eds.): *Classification and Information Processing at the Turn of the Millenium, Procs. of the 23^{rd} Annual Conference of the Gesellschaft für Klassifkation, Bielefeld, 10–12 March, 1999.* Springer, Berlin, 18–31.

FELSENSTEIN, J. (1988): Phylogenies from Molecular Sequences: Inference and Reliability. *Annual Review of Genetics, 22, 521–565.*

FIELD, C.M.; AL-AWAR, O.; ROSENBLATT, J. WONG, M.L., ALBERTS, B. and MITCHISON, T.J. (1996): A Purified Drosophila Septin Complex Forms Filaments and Exhibits GTPase Activity. *Journal of Cell Biology, 133, 605–616.*

FIELD, C.M. and KELLOGG, D. (1999): Septins: Cytoskeletal Polymers or Signalling GTPases? *Cell Biology, 9, 387–394.*

FITCH, W.M. and MARGOLIASH, E. (1967): Construction of Phylogenetic Trees. *Science, 155, 279–284.*

FITCH, W.M. (1971): Toward Defining the Course of Evolution: Minimum Change for a Specified Tree Topology. *Systematic Zoology, 20, 406–416.*

FRAZIER, J.A.; WONG, M.L.; LONGTINE, M.S.; PRINGLE, J.R.; MANN, M.; MITCHISON, T.J. and FIELD, C. (1998): Polymerization of Purified Yeast Septins: Evidence that Organized Filament Arrays may not be Required for Septin Function. *Journal of Cell Biology, 143, 737–749.*

GOLDMAN, N. (1996): Phylogenetic Estimation. In: M.J. Bishop and C.J. Rawlings (Eds.): *DNA and Protein Sequence-Analysis.* IRL Press, Oxford, 287–312.

GUIMARÃES, G. and URFER, W. (2002): Self-Organizing Maps and its Applications in Sleep Apnea Research and Molecular Genetics. In: O. Opitz and M. Schwaiger (Eds.): *Exploratory Data Analysis in Empirical Research. Studies in Classification, Data Analysis and Knowledge Organization.* Springer, Heidelberg, 332–345.

HARTWELL, L.H. (1971): Genetic Controll of the Cell Division Cycle in Yeast. IV. Genes Controlling Bud Emergence and Cytokinesis. *Experimental Cell Research, 69, 265–276.*

HUELSENBECK, J.P. and HILLIS, D.M. (1993): Success of Phylogenetic Methods in the Four-Taxon Case. *Systematic Biology, 42, 247.*

KARTMANN , B. and ROTH, D. (2001): Novel Roles for Mammalian Septins: From Vesicle Trafficking to Oncogenesis. *Journal of Cell Science, 114, 839–844.*

KOHONEN, T. (1982): Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics, 43, 59–69.*

KOHONEN, T. (2001): *Self-Organizing Maps.* 3rd ed., Springer, New York.

LEIPE, D.D.; WOLF, Y.I.; KOONIN, E.V. and ARAVIND, L. (2002): Classification and Evolution of P-loop GTPases and Related ATPases. *Journal of Molecular Biology, 317, 41–72.*

LONGTINE, M.S.; DE MARINI, D.J.; VALENCIK, M.L.; AL-AWAR, O.S.; FARES, H.; DE VIRIGILIO, C. and PRINGLE, J.R. (1996): The Septins: Roles in Cytokinesis and Other Processes. *Current Opinion in Cell Biology, 8, 106–119.*

MÁRKUS, L.; BERKE, O.; URFER, W. and KOVÁCS, J. (1999): Spatial Prediction of the Intensity of Latent Effects Governing Hydrological Phenomena. *Environmetrics, 10, 663–654.*

MOMANY, M.; ZHAO, J.; LINDSEY, R. and WESTFALL, P.J. (2000): Characterization of the Aspergillus Nidulans Septin (asp) Gene Family. *Genetics, 157, 969–977.*

SAITOU, N. and NEI, M. (1987): The Neighbor-Joining Method: a new Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution, 4, 406–425.*

SANKOFF, D. and CEDERGREN, R.J. (1983): Simultaneous Comparison of Three or More Sequences Related by a Tree. In: Sankoff, D. and Kruskal, J.B. (Eds.): *Time Warps, String Edits, and Macromolecules: theTheory and Practice of Sequence Comparison.* Addison-Wesley, 253–264.

SARASTE, M.; SIBBALD, P.R. and WITTINGHOFER, A. (1990): The P-Loop – a Common Motif in ATP- and GTP-Binding Proteins. *Trends in Biochemical Science, 15, 430–434.*

SOKAL, R.R. and MICHENER, C.D. (1958): A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Scientific Bulletin, 28, 1409–1483.*

STRIMMER, K. and von HAESELER, A. (1996): Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. *Molecular Biology and Evolution, 13, 964–969.*