Published online 6 August 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 18* ***5959–5968***
doi:10.1093/nar/gkp634

Detecting species-site dependencies in large multiple sequence alignments

Roland Schwarz[1,2], Philipp N. Seibel[2], Sven Rahmann[3], Christoph Schoen[1], Mirja Huenerberg[4], Clemens Müller-Reible[4], Thomas Dandekar[2], Rachel Karchin[5], Jörg Schultz[2] and Tobias Müller[2,*]

[1]Institute of Hygiene and Microbiology, Josef-Schneider-Strasse 2/E1, 97080, [2]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, [3]Bioinformatics for High-Throughput Technologies at the Chair of Algorithm Engineering (Ls11), Computer Science Department, TU Dortmund, 44221 Dortmund, [4]Institute of Human Genetics, Biocenter, University of Würzburg, Würzburg, Germany and [5]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21204, USA

Received May 5, 2009; Revised June 29, 2009; Accepted July 16, 2009

ABSTRACT

Multiple sequence alignments (MSAs) are one of the most important sources of information in sequence analysis. Many methods have been proposed to detect, extract and visualize their most significant properties. To the same extent that site-specific methods like sequence logos successfully visualize site conservations and sequence-based methods like clustering approaches detect relationships between sequences, both types of methods fail at revealing informational elements of MSAs at the level of sequence–site interactions, i.e. finding clusters of sequences and sites responsible for their clustering, which together account for a high fraction of the overall information of the MSA. To fill this gap, we present here a method that combines the Fisher score-based embedding of sequences from a profile hidden Markov model (pHMM) with correspondence analysis. This method is capable of detecting and visualizing group-specific or conflicting signals in an MSA and allows for a detailed explorative investigation of alignments of any size tractable by pHMMs. Applications of our methods are exemplified on an alignment of the Neisseria surface antigen LP2086, where it is used to detect sites of recombinatory horizontal gene transfer and on the vitamin K epoxide reductase family to distinguish between evolutionary and functional signals.

INTRODUCTION

Multiple sequence alignments (MSAs) are high dimensional discrete datasets, which play a prominent role in bioinformatics. They are typically involved in the functional classification of proteins and phylogenetic reconstruction of evolutionary trees, for example. In general, there are two aspects of MSAs; analyses are mostly either species- or site focused. Species-driven approaches usually aim at the relationship between sequences, averaging over the alignment columns. Methods for phylogenetic reconstruction as well as general sequence clustering methods are examples, and make (amongst other things) use of distance measures to impose an hierarchy on the species in an alignment. This allows for the detection of closely related species, functional clusters and the reconstruction of gene trees or species trees. Site-driven analyses in contrast put more emphasis on sequence content, looking for specific sequence motifs, conservation profiles, areas with characteristic biochemical properties like hydrophobicity or transmembrane regions, thereby averaging over the sequences or focusing on their conserved regions.

A combination of both types of analyses of an (correctly aligned) MSA helps to distinguish functionally conserved from variable sites, detect clusters of sequences and find sites responsible for a certain splitting of sequence groups. This integration can finally lead to an understanding of the functional evolution of sequences, as tree splits or cluster breaks can be annotated with the associated autapomorphies [an autapomorphy is a trait characteristic for a terminal group in a phylogenetic tree

*To whom correspondence should be addressed. Tel: +49 931 318 4563; Fax: +49 931 318 4552; Email: tobias.mueller@biozentrum.uni-wuerzburg.de

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

(a monophyletic group), i.e. a property that is shared by only the members of the group, but not by any other taxa]. Due to the complexity of MSAs of realistic size, thorough analyses require expert knowledge, are tedious, time consuming and error-prone.

Traditionally, first view analyses are done in alignment editors/aligners like SEAVIEW (1), CLUSTAL_X (2), Jalview (3) or 4SALE (4). Amino acids are usually colored with respect to their biochemical and physical properties and conservation bars are aligned to the MSA to get a column-based summary. A better graphical representation of the degree of conservation can be achieved by sequence logos (5), which additionally visualize the entropy of the site distributions. RNA logos also include horizontal dependencies in RNA sequences, defined by their respective secondary structure (6,7). With the arrival of hidden Markov model (HMM) (8–10) in sequence analysis, HMM logos were introduced presenting entropy terms based on estimated HMM parameters like emission, insertion and deletion probabilities (11,12).

These site-focused methods provide an abstract summary of the sequence variability in an alignment, but usually do not allow for the detection of sequence clusters and fail at representing long sequences adequately. Apart from character-based methods, clustering of sequences is either done indirectly, via an interposed distance measure as in the case of phylogeny, or requires a meaningful way to embed sequences into a real-valued vector space, something which cannot be achieved trivially. Given such an embedding, standard dimension reduction techniques like principal component analysis (PCA) or classical multidimensional scaling (MDS) could be applied. Casari *et al.* (13) introduced a method for dimension reduction on MSAs, which was later implemented in the Jalview application (3). The algorithm is based on a simple mapping of sequences to binary vectors, not including gaps, and applies PCA to the binary sequence data.

Our method captures both horizontal and vertical information by combining an improved embedding of sequences including gaps with a site-specific annotation of sequence clusters. Instead of mapping the sequence data to a binary vector, we apply an HMM-based embedding using a vector of sufficient statistics for the emission probabilities instead of the Fisher scores (14–16). We apply correspondence analysis (CA) (17) to the embedded sequences and sites, elaborating on the association between both data and visualizing clusters of sites and sequences in one joint plot. Dimension reduction is done the usual way, preserving as much information as possible in the lower dimensional representation. Selection of the axes allows for a precise investigation of different signals in the alignment, as shown in studies on the Neisseria factor H binding protein and the vitamin K epoxide reductase family.

MATERIALS AND METHODS

Embedding

Molecular sequences are typically represented by strings over an alphabet of either 4 or 20 characters. In order to apply numerical methods on these kinds of data, a sensible embedding into $\mathbb{R}^n$ has to be found. Fisher scores are derived from the posterior probabilities of a fitted HMM and are known to be a sufficient statistic for the fitted HMM parameters (15,18).

Fisher scores are the derivative of the log-likelihood of an HMM with respect to all parameters of the HMM, namely emission and transition probabilities, evaluated for each datum, i.e. for each sequence. To be more precise, the Fisher score vector F_i for the i-th sequence S_i is $F_i = \nabla_\Theta \log(P[S_i|\Theta])$, where Θ denotes the vector of HMM parameters. They therefore represent a site-specific fixed-length embedding that directly encodes emission, insertion and deletion events.

Intuitively spoken, a Fisher score of an HMM parameter describes the slope of the likelihood for the given data (the given sequence) with respect to this parameter. This can be seen as the degree of influence the datum has on the parameter in an optimization context, or the degree of surprise encountering the given amino acid/nucleotide/indel at that specific alignment position. For a precise description of the computational details of the Fisher score calculation, see refs (15,19).

Correspondence analysis (CA)

CA is an ordination method originally created for count data in two-way contingency tables and rooted in ecology and community analysis (17). In contrast to other ordination methods built around singular value decomposition (SVD), CA performs its ordination simultaneously on column and row scores. It superimposes the results in one joint plot, thus painting an usually 2D picture of dependencies between data points and its most significant factors.

Pre-processing. For technical reasons, the $n \times m$ data matrix $F = (f_{ij})$ is first made positive by adding a constant to each entry. It is then normalized by dividing the matrix entries by its respective row and column sums $h_{ij} = f_{ij}/\sqrt{f_{i\cdot}f_{\cdot j}}$, resulting in the normalized data matrix H. In matrix notation this may be written as $H = S^{-1/2} XC^{-1/2}$, where $S^{-1/2}$ and $C^{-1/2}$ are diagonal matrices containing the reciprocals of the square root of the row and column marginal totals.

SVD. SVD is a factorization of a real or complex matrix $A \in M(m \times n;\ \mathbb{K})$ of the form $A = U\Sigma V^*$ where U is a $m \times n$ unitary matrix over a field $\mathbb{K}$, Σ is $n \times n$ positive semidefinite diagonal matrix and V^* denotes the conjugate transpose of V, an $n \times n$ unitary matrix over $\mathbb{K}$. Σ contains the singular values, whereas the columns of U and V^* are the left- and right-singular vectors for the corresponding singular values (20).

A lower dimensional representation of the data is generated by ordering the singular values by size and taking the first n' singular values. The loss of information is described in terms of the proportion of the sum of squares of singular values $\sum_{i=1}^{n'} \Sigma_{ii}^2$ used (total inertia). In the CA context, the total inertia is proportional to the value of the χ^2 statistic, and thus to the degree of association in the data (21).

Post-processing. After SVD, the row U and column V scores are usually rescaled via $X_i = U_i\sqrt{f_{\cdot\cdot}/f_{i\cdot}}$ and $Y_i = U_i\sqrt{f_{\cdot\cdot}/f_{i\cdot}}$, to obtain the optimal or canonical row (X) and column (Y) scores. Depending on the implementation of the CA algorithm, these are afterwards further scaled by their corresponding singular values (17).

Interpretation. The selected component axes are then plotted in usually 2D scatterplots. The Euclidean projection of both site and species points in the new space approximates their χ^2 distances as closely as possible. Proximity of points in the CA biplot therefore corresponds to dependencies between items. Furthermore, points are projected such that the further away a point is from the origin, the higher its contribution to the χ^2 statistic. Positive associations lie on the same side of the plot, whereas negative associations lie on the opposite sides. For a more detailed explanation, see ref. (17). Please note that the addition of a positive constant to the original data matrix does not change the proportions of the new coordinate system but implies a rescaling of the result.

Sequence analysis

Neisseria meningitidis factor H binding protein. Sequences for the LP2086 and VKOR studies were aligned using Muscle (22) (Supplementary Data). The distance matrix for the LP2086 alignment was calculated using *ProfDist* (23,24) applying the VT substitution matrix (25). The distance matrix was further analyzed and visualized by *SplitsTree*'s split decomposition method (26).

Vitamin K epoxide reductase family. Vertebrate sequences were extracted from ENSEMBL using the human VKORC1 and VKORC1L1 proteins as query in a blastp search (27). The ENSEMBL identifiers are:

L1_Human	ENSP00000353998	L1_Pan	ENSPTRP00000047967
L1_Macaca	ENSMMUP00000027960	L1_Rat	ENSRNOP00000024691
L1_Mouse	ENSMUSP00000073601	L1_Monodelphis	ENSMODP00000008083
L1_Xenopus	ENSXETP00000022171	L1_Danio	ENSDARP00000064087
L1_Oryzias	ENSORLP00000014137	L1_Fugu	ENSTRUP00000017074
C1_Human	ENSP00000378426	C1_Pongo	ENSPPYP00000008254
C1_Cat	ENSFCAP00000002398	C1_Horse	ENSECAP00000021915
C1_Dog	ENSCAFP00000024701	C1_Cow	ENSBTAP00000000519
C1_Rat	ENSRNOP00000026347	C1_Mouse	ENSMUSP00000033074
C1_Fugu	ENSTRUP00000027115	C1_Tetraodon	ENSTNIP00000018260

The *Ciona savigny* homologue was identified only in genomic sequences. The protein sequence was predicted using gene-wise (28) and the human VKORC1 protein as template. The alignment was calculated using Muscle (22) and manually optimized (Supplementary Data). The phylogenetic tree for the VKOR example was calculated with *proml* of the PHYLIP package (29) and 100 bootstrap replicates. Ancient sequences were reconstructed by *codeml* of the PAML package (30).

RESULTS

The method we propose here is a novel approach to an explorative analysis and visualization of MSAs. The goal of our method is the detection and depiction of major signals in alignments, ordered by their importance, co-clustering of sequences and sites and resolution of contradictory signals, i.e. different parts of the alignments vote for a different clustering.

The approach comprises three separate steps: (i) the embedding of sequence data into a real valued very high dimensional vector space, (ii) the simultaneous dimension reduction and ordination of both rows and columns of the data matrix (the alignment) and (iii) a biplot visualization of the canonical row and columns scores.

The result is a lower dimensional representation of sequences and sites, which can be analyzed by (two-, or three dimensional) scatterplots, comparable but not identical with the result of classical dimension reduction techniques like PCA, applied to both sequences and sites. In contrast to traditional dimension reduction methods, the sequences are co-clustered to their defining sites and vice versa. In this representation, the sites responsible for a cluster of sequences come to lie close to the sequences.

Embedding (i) is achieved via the Fisher score representation of HMM parameters (14,15,18). Therefore we start by training a profile HMM (pHMM), (9,10) on the previously aligned sequences. The Fisher scores are then computed as the vector of derivatives of the log-likelihood of each training sequence with respect to the emission probabilities of the HMM (see 'Material and Methods' section). The sequence is thus transformed in a meaningful way into a vector of real-valued numerical values for the following ordination step.

Steps (ii) and (iii) are done via direct application of CA to the derived data matrix of Fisher scores. CA is a method originating from ecology and designed for the analysis of two-way contingency tables (17). It is capable of performing simultaneous ordination on both rows and columns of a data matrix (often referred to as species and sites in ecology, a nomenclature which also fits well in sequence analysis) and has also been shown to be of use for continuous datasets in the context of microarray analysis (31). In principle, it can be thought of as an oriented MDS on χ^2 distance matrices computed from both sides of a data matrix, which is jointly plotted. In the CA, each axis is a weighted linear combination of the Fisher scores of the data vectors, i.e. of the existent (and due to the way the Fisher scores are generated also non-existent) nucleotides/residues in the alignment. CA is a co-clustering of sequences and sites, where conditionally independent signals are projected onto the component axes.

Therefore, in a phylogenetic context, one would expect that the first component axis corresponds to the branch of a phylogenetic tree which discriminates most between the most different sequence groups. Typically this refers to the longest branch of the tree. This means that the two major phylogenetic sequence groups are expected to lie consistently on one side of the first component axis, or the other, respectively. Other long-branched subgroups of the tree are then likely to be found in higher order component axes. The co-clustered sites are major candidates for the autapomorphies defining the split. In the same manner, alignments can be decomposed, even when a well-supported phylogenetic tree cannot be constructed, either due to contradictory signals within the alignment or different evolutionary histories. In summary, our proposed method yields a complete decomposition of the considered MSA. In particular, it visualizes information content and species-site dependencies with respect to a given sequence family, modeled by the underlying pHMM.

Example on an artificial dataset

To illustrate the concept of our proposed method, we created an artificial DNA MSA (Figure 1a) of four sequences. The main split of the associated cluster tree (Figure 1b) distinguishes the sequences 1 and 2 from 3 and 4. Given the first split, split II distinguishes between sequences 1 and 2 and split III distinguishes between sequences 3 and 4.

Application of our method illustrates how it is able to recover the sequence groups and the nucleotide replacements responsible for the grouping. The procedure decomposed the alignment into a 3D space, without loss of information. Figure 1c and d are CA plots of the MSA showing the first three component axes (1 versus 2, and 3 versus itself). In Figure 1c, the first component axis corresponds to the main split of the cluster tree and separates sequences 1 and 2 from sequences 3 and 4, thereby indicating the sites responsible for this split, i.e. G and C versus both Ts at position 3 and 4. The second component axis explains the (conditional) split between sequences 1 and 2 identified by an A or T at position 10. The last conditional split to be explained is the one separating sequences 3 and 4. This is shown in Figure 1d, the third component axis, which identifies the differences at position 16 (G versus C) as being responsible for the split.

Neisseria meningitidis factor H binding protein

To validate our method on a biological example we chose the *N. meningitidis* factor H binding protein (fHBP), also termed lipoprotein 2086 and GNA1870, which has become a prominent target in the development of a novel vaccine against serogroup B meningococci (32–34). This alignment seemed especially suitable for evaluating our method, as there have been conflicting reports about how many distinct sequence variants can be found within the sequence cluster (32,33). We based our analysis on an extended alignment consisting of 114 (47 distinct) sequences from the Genbank database, including the 64 (21 distinct) sequences used by Fletcher *et al.* (33). We skipped the initial clustering step proposed by Fletcher and co-workers and worked directly on the complete alignment of 114 sequences, each 263 amino acids long. Embedding and ordination took ~30 s on a standard desktop computer. Distance-based phylogenetic analyses carried out by the Fletcher group showed a clear clustering of the sequences into two separate subfamilies, each with several further sub-clusters. The authors concluded from these findings that the sequence family consists of two major sequence variants (called subfamily A and B) and recommended representatives of those two variants to be used for vaccine design.

On the contrary, Masignani *et al.* (32) reported at least three major sequence clusters and consequently recommended to use representatives of all three clusters to be included for vaccine design against serogroup B meningococcal disease.

Application of our method of combined embedding and ordination of the sequence alignment resulted in a 46-dimensional representation of the data matrix comprising 5610 Fisher score columns. The distribution of the cumulative contribution of the axis to the χ^2 statistic showed typical exponential behavior, where seven axes were sufficient to explain >50% of the total inertia. Visual inspection of the major contributing axes showed that axes 1–3 were prominent candidates for the detection of the major sequence clusters (Figure 2c), where the signal on axis 2 was mainly due to single nucleotide polymorphisms in otherwise highly conserved positions.

Investigation of the scatterplot showed clustering of the sequences in four separate groups (Figure 2c). Axis 1 separated the Fletcher subfamily A (left, negative half-plane) from subfamily B (right, positive half-plane) without error. From the co-clustered sites it could be seen that major blocks of conserved sites within the respective groups, ranging from alignment position 106–261, were mainly responsible for the observed grouping (Figure 2b, right side).

To compare our results with classical methods, we computed a matrix of evolutionary distances between all 47 unique sequences, which was then visualized as an evolutionary network using split decomposition (26). The main cluster, as found by Fletcher *et al.* (33) and our analysis of component axis 1, was also recovered in the evolutionary network (Figure 2a). These findings strongly suggest that the evolutionary split which lead to development of subfamilies A and B must have happened early in the history of this protein.

Remarkably, axis 3 divided both subfamilies A and B into two sub-clusters (A1, A2 and B1, B2, respectively). When we investigated the most prominent representatives of these groups (i.e. the ones closest to the borders of the plot), the co-clustered sites showed that these groups contain identical sequence elements (positions 37–69), including a three-residue long lys-asp-asn insertion between alignment position 67 and 69.

This indicates that if the development of subfamilies A and B was prior to the emergence of the second split, clusters 1 and 2 have developed within subfamily A (Figure 2a), and parts of the sequence has afterwards been transferred to members of subfamily B by means of an horizontal gene transfer (HGT)/recombination event. This uncertainty in the evolutionary hierarchy between the sequences is also reflected in the large rectangles contained in the split decomposition visualization of the distance matrix (Figure 2a). This finding was further supported by a PHI test for recombination, which was carried out on the complete alignment (P-value $<1.07 \times 10^{-11}$) (35).

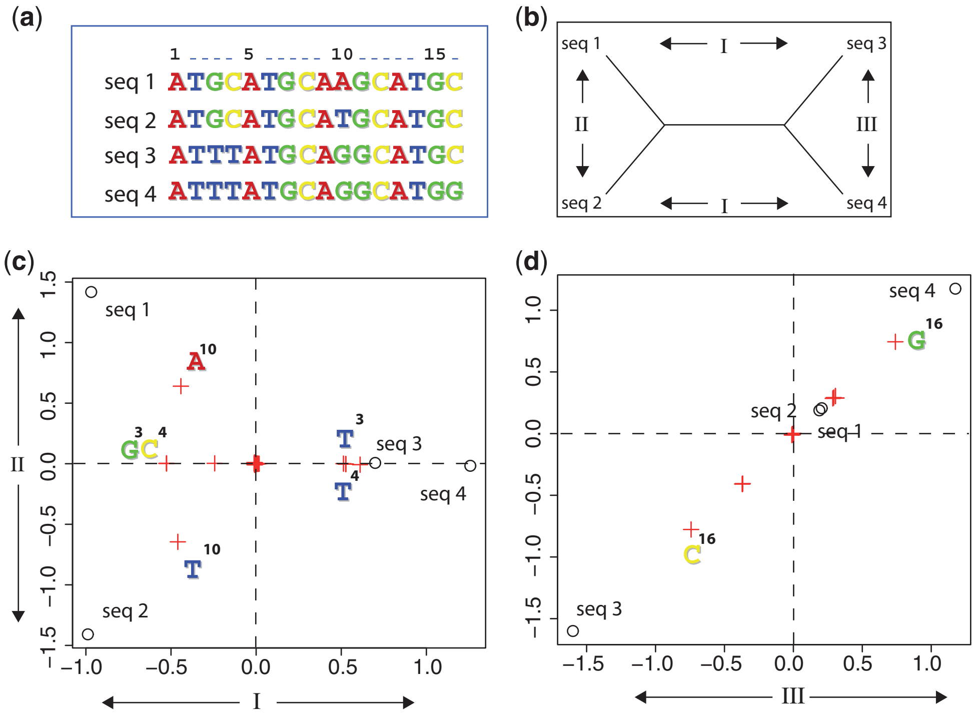

Figure 1. Artificial example: the MSA (**a**) and its cluster tree (**b**) as used in our toy example. The subparts (**c**) and (**d**) are scatterplots of the first three component axes, which together account for 100% of the inertia in the data. The CA plots present sequences (black circles) and sites (red crosses) in an integrated manner. For better interpretation, the most important sites are explicitly shown in the plots with their nucleotide letters and alignment positions. Roman numbers indicate the splits in the cluster tree and the component axes resolving them.

Vitamin K epoxide reductase family

Vitamin K is an essential cofactor for the post-translational γ-glutamyl carboxylation of the vitamin K-dependent proteins such as several coagulation factors, bone proteins, cell growth regulating proteins and others of unknown function (36,37). During the carboxylation vitamin K hydrochinone is converted into vitamin K 2,3-epoxide (38). The recycling reaction of vitamin K epoxide back to the hydrochinone form is catalyzed by the vitamin K epoxide reductase (VKORC1) in the so-called vitamin K cycle (39). VKORC1 is the key protein in this redox reaction and the molecular target of coumarin derivatives, such as warfarin, which act as vitamin K antagonists (40). They reduce coagulation activity by interfering with the vitamin K epoxide reductase. Worldwide, coumarins are used in therapy and prevention of thromboembolic events and also in higher doses for rodent pest control. Mutations in the *VKORC1* gene cause one form of combined deficiency of vitamin K-dependent coagulation factors (VKCFD type 2) as well as resistance or hypersensitivity to warfarin (41,42). The human *VKORC1* gene is localized on chromosome 16 (43) and consists of three exons encoding a 163-amino acid endoplasmic reticulum membrane protein with three or four predicted transmembrane α-helices (44). With the identification of the *VKORC1* gene in 2004 (45,46) a paralog gene was discovered, which is called vitamin K epoxide reductase complex 1-like 1 (*VKORC1L1*) and which is highly conserved over the species. Its physiological function is completely unknown.

Extensive database searches in a wide variety of metazoan genomes and subsequent phylogenetic reconstruction allowed us to time the duplication event to the base of the vertebrates. To identify candidate positions for functional analyses, we built a MSA including both variants over different vertebrates, namely a group of fish species (danio, tetraodon, fugu and oryzias), a group of mammals (macaca, pan, human, pongo, mouse, rat, cow, cat, dog and horse), Monodelphis and Xenopus. Furthermore, the alignment contained the VKOR ortholog of *Ciona savigny*, pre-dating the duplication event. As expected, a first phylogenetic tree revealed two groups, VKORC1 and VKORC1L1, and placed the *C. savigny* sequence as outgroup. It further clearly separated the fish species from the rest in both groups and correctly clustered the subgroups of mammals in contrast to the singletons Chicken, Monodelphis and Xenopus (Figure 3a).

Application of our method to this alignment revealed the following: the first (and most informative) axis separated all species, i.e. all duplicated genes, from the *C. savigny* sequence (data not shown). This corresponds to the longest branch and rootsplit in the phylogenetic tree, but as we were more interested in variation between species with both paralogs present, we did not investigate this further. We expected axis 2 to either separate the C1 from the L1 sequences or the fish from the land animals, in analogy to the phylogenetic tree. Axis 3 in general separated C1 from L1, for all but the C1 fish sequences, which came to lie near the origin. Analysis of

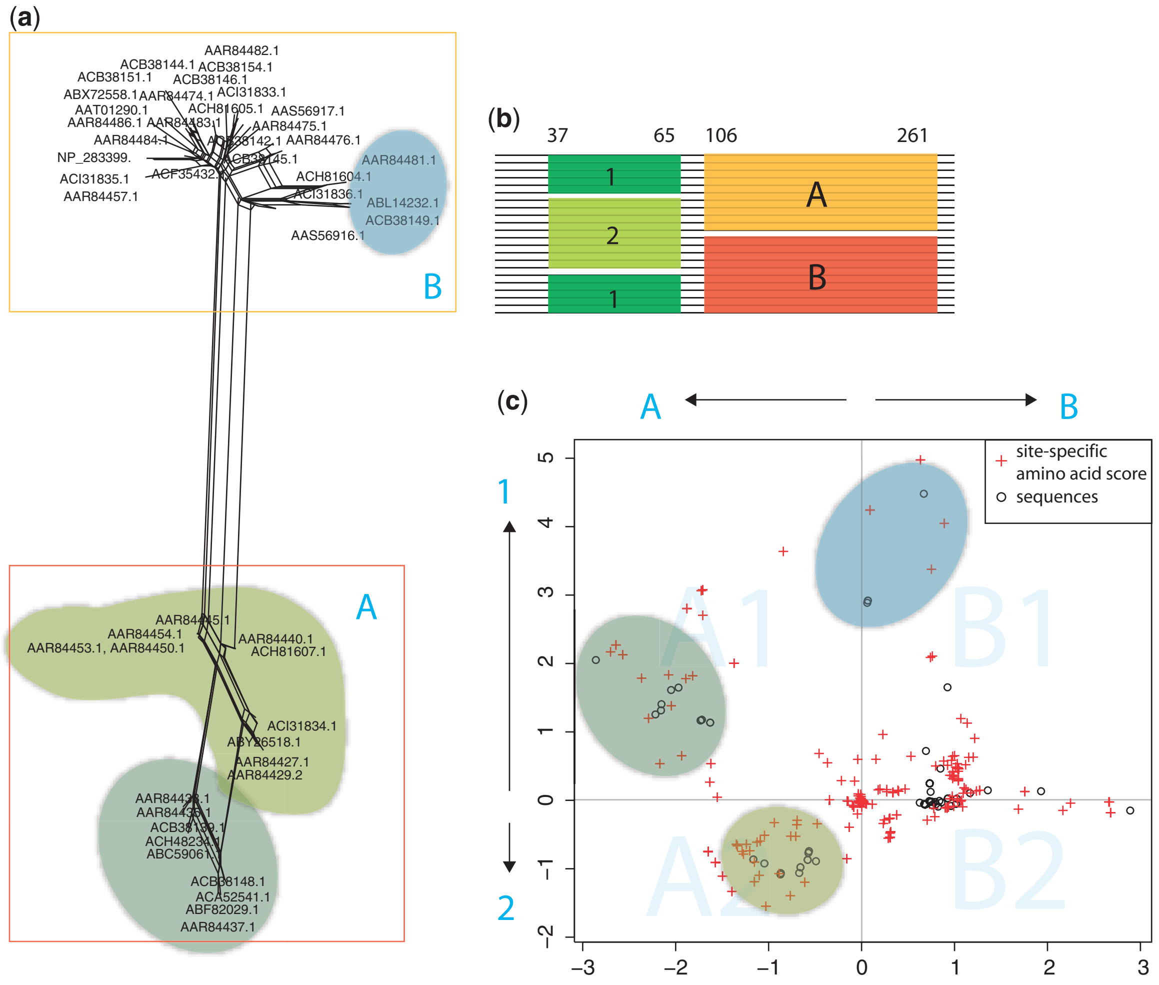

Figure 2. Analysis of the LP2086 sequence family. (**a**) Evolutionary network reconstructed from a distance matrix on 47 unique sequences. Fletcher subfamilies A and B are clearly separated. The further sub-clusters 1 and 2 are marked in color. (**b**) Schematic representation of the complete alignment of 114 LP2086 sequences, where major parts of the alignment (from position 100 onward) have a block structure corresponding to Fletcher subfamilies A and B, a 30 amino acid region in the beginning votes for a different grouping. (**c**) CA plot of component axes 1 and 3. The method groups the relevant clusters, isolating each from the rest, and identifies the relevant sites. The groups are colored in analogy to those in the evolutionary network.

the subsequent axes of the ordination results separated the L1 fish sequences from its main group (axis 5) and showed that the Danio sequence within the C1 fish group was evolutionarily more distant to the other C1 fish (axis 4), as reflected in the phylogenetic tree.

The co-clustered sites showed us that positions which are otherwise completely conserved within the C1 or L1 family were different in the C1 fish sequences. For example, alignment positions 73–77 (marked with yellow dots in Figure 3b) showed a typical EHVL motif for the C1 family and a GSIF sequence for the L1 cluster. The C1 fish sequences in contrast had a QYFV motif (QIFT for Danio) instead. The missing information was caught by axis 2 which separated the C1 fish sequences from all others (for a combined scatterplot of axes 2 and 3, see Figure 3b). In addition, different positions in the alignment were identified, where the fish C1 sequences harbored the same amino acids as the L1 land animal group but differed from the rest of the C1 group. A prominent example is the Warfarin binding motif which is found as a TYA in the C1 non-fish and L1 fish sequences, but as a TYV/TYI/TYL in the C1 fish and L1 non-fish sequences. Reconstruction of ancient sites revealed that this motif evolved in the C1 group only after the split of fishes from the other vertebrates (Figure 3a). Following this observation, we extracted all positions specific for the L1 group and the L1/C1 fish groups, respectively.

To analyze their functional relevance, we mapped these positions onto the transmembrane topology of this protein [Figure 3b, (44)]. Two clusters of these sites reside on the cytoplasmic extensions of the transmembrane helices I and III. Further sites are localized within the transmembrane helix II. Here, the positions were placed regularly on every fourth position (alignment positions 111, 115, 119 and 123, Figure 3b).

With a standard helix turn taking on about every 3.5 amino acids, there seems to be a spatially aligned position, where the sites of this transmembrane helix are specific for the subgroups. Although highly speculative, these findings might suggest the following model of action for

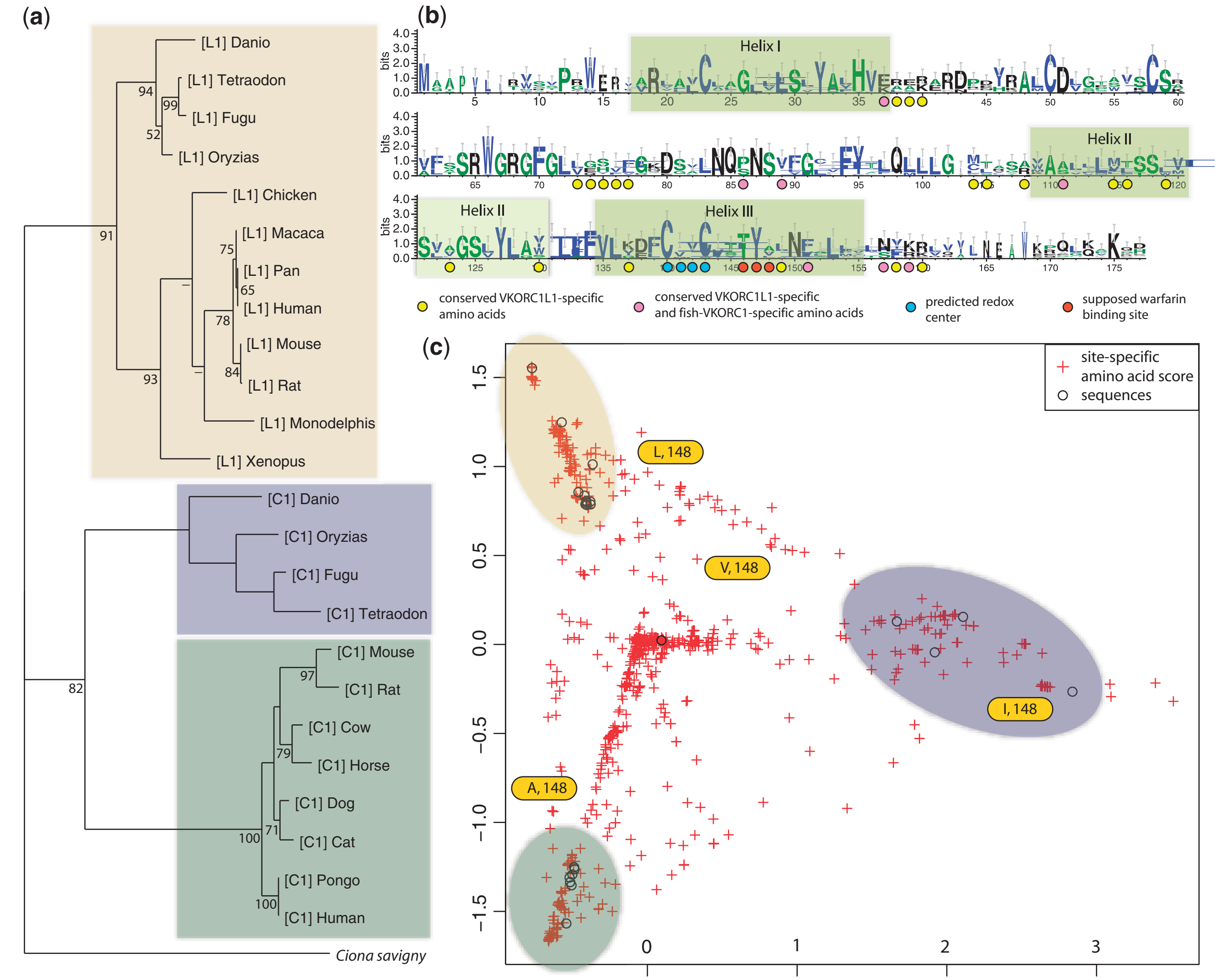

Figure 3. Analysis of the VKOR sequence family. (**a**) Phylogenetic tree of the VKOR protein family. (**b**) Sequence logo of the MSA including the proposed membrane topology of VKORC1 with conserved positions for VKORC1L1 (44). The conserved VKORC1L1-specific amino acids are marked in yellow. Pink-labeled amino acids are specific to VKORC1L1 and to the VKORC1 protein of fish. In the third transmembrane domain, the blue circles symbolize the redox center (CIVC motive) and the supposed warfarin binding site with the TYA motive is highlighted in red. (**c**) Scatterplot of the second and third principal factors. Sequences are depicted as black circles, sites as red crosses. Closeness of sequences and sites in the plot shows strength of association.

this family of transmembrane proteins. First, a substrate, differing between the C1 and L1 subfamilies, is bound by the cytoplasmic extensions of helices I and III. Possibly, a further region in the first, large cytoplasmic loop (position 73–77 in Figure 3b) assists in substrate recognition. Second, the substrate is channeled into the membrane along one site of transmembrane helix II. Finally, it is presented to the catalytic center built by the CIVC motif residing in helix III (blue dots, Figure 3b).

DISCUSSION

Recent advances in genome sequencing technology have lead to a noticeable shift in focus toward methods dealing with contig- or genome-sized sequences, be it for contig assembly or phylogenomics. Nevertheless, accurately reconstructed MSAs on the gene or protein level are still of major importance. Most tools or algorithms introduced in this context are dedicated to a specific task like the reconstruction of phylogenetic trees, transmembrane prediction or conservation profiling.

The method we propose here is different in that it is a method for the explorative unsupervised analysis of MSAs. It decomposes the alignment into its major signals and co-clusters sequences and sites, thereby simultaneously finding sequence groups and the sites responsible for their grouping. The probabilistic model (pHMM) used to describe the alignment is a known and approved method for sequence modeling (10) and due to their nature the Fisher score embedding is advantageous to other embeddings proposed and applied before (3,13). These advantages include the possibility to directly model and encode transition probabilities of the pHMM and thereby insertions and deletions in the alignment. Further, apart from a pure probabilistic representation of the alignment itself, the HMM fitting process allows integration of prior knowledge about amino acid distributions. Biologically meaningful priors can be derived, e.g. via Dirichlet mixtures (47,48) or from log-odds-based substitution matrices (49). These incorporate the desired biological signal into the pHMM, giving, e.g. amino acid positions with similar chemical or physical properties, more similar probabilities than obtained from the alignment alone.

Fisher scores are known to be 'sufficient statistics' for the underlying HMM parameters, i.e. they contain all available information about the parameters (14). In contrast to a direct embedding via the HMM scores or site probabilities, they do not suffer from the effect that highly divergent, but from the HMM's perspective equally probable sequences receive the same representation. This would project those unrelated sequences close to each other during the ordination step. Additionally, Fisher scores are a fixed-length representation of the original sequences, thus preventing length-driven biases in the analyses. Computational problems due to the high dimensionality of the Fisher score representation itself can be circumvented by application of the economy-sized SVD variant. The computational complexity of the Fisher score calculation is similar to the forward–backward algorithm [$O(N^2T)$ for N states and sequence length T].

Even though our proposed method of ordination (CA) was originally designed for two-way contingency tables (17), it has been shown earlier that the method is very suitable for the analysis of continuous datasets, in which dependencies between rows and columns of a data matrix are of interest (31).

We compared our method to a standard approach of ordination with an Euclidean metric (e.g. PCA). Representatives are, for example, the SeqSpace and Jalview programs (3,13,50), although these tools additionally suffer from the inexpressiveness of the binary embedding employed. For a fair competition, we compared our CA decomposition to classical PCA on the same dataset, in both cases embedded via Fisher scores, and found CA to be more sensitive toward biological signals. For example, PCA analysis of the LP2086 dataset moved sequences ACB38144.1 and ACI31835.1 (close to the blue HGT candidate group in Figure 2c) even though they do not share the 30 amino acid region characteristic for sequences of that cluster (Figure 2b). In the original CA ordination, they clearly separate from the other sequences of their cluster on the x-axis (the two points on the far right side of Figure 2c), but show no grouping with the HGT candidates. Similar effects were found in other regions of the sequence and in the VKOR example (data not shown). It seems that CA profits from application of the χ^2 distance in that it focuses on sequence–site associations rather than simple one-way Euclidean ordination. We finally also directly loaded our datasets into Jalview, but as the software is missing the ordination of sites in the alignment, no functional annotation of sequence clusters could be made. The SeqSpace software, which is supposed to also cluster the sites, was not available anymore at the time of this writing.

The advantages of detecting associations in terms of the χ^2 distance become apparent in the fHBP example. Neither sequence-based nor site-based methods are on their own able to detect any recombination event. Phylogenetic algorithms average over the length of the alignment, rightfully discarding the subtle 30 amino acid transfer region in the beginning of the alignment. The HGT never shows in the tree, it can be suspected from the evolutionary network, but due to the short length and the low number of representatives carrying the motif, the signal is only weakly reflected in the distance matrix and therefore in the split decomposition. Conservation profiles like sequence logos or clustering procedures on sites would not reveal the HGT either, which can only be identified by detection of incompatible sites (35), i.e. sites for which contradicting sequence clusters can be built. Our method was able to resolve the recombinated group and identify the responsible sites. It allows for an explorative analysis of the MSA without focusing on any specific type of signal, e.g. phylogenetic signals or HGT alone. It is important to note that this is by no means a test for recombination nor a method to thoroughly find all possible sites of HGT within an alignment, but it can provide an unbiased and structured view on an MSA from different perspectives.

Studying the VKOR protein family again showed how major phylogenetic signals appear on one of the first axes in the ordination, like separation of the *C. savignyi* outgroup. But it is also a good indication of how interesting features of the alignment are completely missed by sequence-based methods, like the phylogenetic tree, or site-base methods, like the depicted sequence logo alone. The co-clustering of species and sites, i.e. the identification of associations between the two, bring insight into the dependencies and—maybe—functional relations, between sequences in the alignment, thereby annotating them with the necessary sequence features. It showed us for example, that in contrast to the L1 fish sequences, the C1 fish sequences do not share the typical C1-L1 site differences of the other groups and identified the positions where those sequences differed. Recovering this tiny signal covered by the large phylogenetic trend would not be possible by methods considering complete sequences, as in the calculation of phylogenetic trees.

From these findings we are convinced that the method proposed here provides researchers with a new and unique way to analyze MSAs. Our method provides a structured decomposition of an alignment and depiction of its information content with increasing granularity. The modularity of the approach allows for a variety of statistical methods applicable to high-dimensional datasets to be used. Its explorative nature can give rise to hypotheses which might then be validated by, for example, statistical tests. On the modeling side, future work might extend the algorithm to include combined sequence structure alignments suitable for analysis of RNA sequences. In general, all types of sequential data (DNA, RNA and protein sequences) are in principle suitable for such an analysis, provided they can be modeled in a probabilistic fashion via, for example, an HMM and from which Fisher scores can be derived.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The software implementing the above described method is available on request from the authors.

FUNDING

DFG (TR34 and SPP1316); BMBF (PathoGenoMik-Plus Grant 0313801A).

Conflict of interest statement. None declared.

REFERENCES

1. Galtier,N., Gouy,M. and Gautier,C. (1996) SEAVIEW and PHYLO-WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–548.
2. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
3. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
4. Seibel,P.N., Müller,T., Dandekar,T. and Wolf,M. (2008) Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Res. Notes*, **1**, 91.
5. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
6. Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci.*, **13**, 583–586.
7. Chang,T.-H., Horng,J.-T. and Huang,H.-D. (2008) RNALogo: a new approach to display structural RNA alignment. *Nucleic Acids Res.*, **36**, W91–W96.
8. Churchill,G.A. (1992) Hidden Markov-chains and the analysis of genome structure. *Comput. Chem.*, **16**, 107–115.
9. Krogh,A., Brown,M., Mian,I.S., Sjölander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
10. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
11. Schuster-Böckler,B., Schultz,J. and Rahmann,S. (2004) HMM Logos for visualization of protein families. *BMC Bioinformatics*, **5**, 7.
12. Hughey,R., Karplus,K. and Krogh,A. (2003) SAM Sequence alignment and modeling software system. *Technical Report UCSC-CRL-99-11*. Update for SAM Version 3.4, Baskin School of Engineering, University of California, Santa Cruz.
13. Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.
14. Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **7**, 95–114.
15. Jaakkola,T., Diekhans,M. and Haussler,D. (1999) Using the Fisher kernel method to detect remote protein homologies. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 149–158.
16. Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
17. Legendre,P. and Legendre,L. (1998) *Numerical Ecology*. Elsevier.
18. Jaakkola,T.S. and Haussler,D. (1998) Exploiting generative models in discriminative classifiers. *Adv. Neural Inf. Process. Syst.*, **11**, 487–493.
19. Shawe-Taylor,J. and Cristianini,N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press.
20. Golub,G.H. and vanLoan,C.F. (1996) *Matrix Computations*. The Johns Hopkins University Press.
21. Weller,S.C. and Romney,A.K. (1990) *Metric Scaling—Correspondence Analysis*. SAGE.
22. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
23. Friedrich,J., Dandekar,T., Wolf,M. and Müller,T. (2005) ProfDist: a tool for the construction of large phylogenetic trees based on profile distances. *Bioinformatics*, **21**, 2108–2109.
24. Wolf,M., Ruderisch,B., Dandekar,T., Schultz,J. and Müller,T. (2008) ProfDistS: (profile-) distance based phylogeny on sequence–structure alignments. *Bioinformatics*, **24**, 2401–2402.
25. Müller,T. and Vingron,M. (2000) Modeling amino acid replacement. *J. Comput. Biol.*, **7**, 761–776.
26. Huson,D.H. and Bryant,D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, **23**, 254–267.
27. Altschul,S.F., Madden,T.L., Schffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
28. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and Genomewise. *Genome Res.*, **14**, 988–995.
29. Felsenstein,J. (2005) *PHYLIP (Phylogeny Inference Package) version 3.6*, (distributed by the author). Department of Genome Sciences, University of Washington, Seattle.
30. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
31. Fellenberg,K., Hauser,N.C., Brors,B., Neutzner,A., Hoheisel,J.D. and Vingron,M. (2001) Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 10781–10786.
32. Masignani,V., Comanducci,M., Giuliani,M.M., Bambini,S., Adu-Bobie,J., Arico,B., Brunelli,B., Pieri,A., Santini,L., Savino,S. *et al.* (2003) Vaccination against Neisseria meningitidis using three variants of the lipoprotein GNA1870. *J. Exp. Med.*, **197**, 789–799.
33. Fletcher,L.D., Bernfield,L., Barniak,V., Farley,J.E., Howell,A., Knauf,M., Ooi,P., Smith,R.P., Weise,P., Wetherell,M. *et al.* (2004) Vaccine potential of the Neisseria meningitidis 2086 lipoprotein. *Infect. Immun.*, **72**, 2088–2100.
34. Giuliani,M.M., Adu-Bobie,J., Comanducci,M., Aric,B., Savino,S., Santini,L., Brunelli,B., Bambini,S., Biolchi,A., Capecchi,B. *et al.* (2006) A universal vaccine for serogroup B meningococcus. *Proc. Natl Acad. Sci. USA*, **103**, 10834–10839.
35. Bruen,T.C., Philippe,H. and Bryant,D. (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics*, **172**, 2665–2681.
36. Shearer,M.J. (2000) Role of vitamin K and Gla proteins in the pathophysiology of osteoporosis and vascular calcification. *Curr. Opin. Clin. Nutr. Metab. Care*, **3**, 433–438.
37. Saxena,S.P., Israels,E.D. and Israels,L.G. (2001) Novel vitamin K-dependent pathways regulating cell survival. *Apoptosis*, **6**, 57–68.
38. Furie,B., Bouchard,B.A. and Furie,B.C. (1999) Vitamin K-dependent biosynthesis of gamma-carboxyglutamic acid. *Blood*, **93**, 1798–1808.
39. Wajih,N., Hutson,S.M. and Wallin,R. (2007) Disulfide-dependent protein folding is linked to operation of the vitamin K cycle in the endoplasmic reticulum. A protein disulfide isomerase-VKORC1 redox enzyme complex appears to be responsible for vitamin K1 2,3-epoxide reduction. *J. Biol. Chem.*, **282**, 2626–2635.
40. Cain,D., Hutson,S.M. and Wallin,R. (1998) Warfarin resistance is associated with a protein component of the vitamin K 2,3-epoxide reductase enzyme complex in rat liver. *Thromb. Haemost.*, **80**, 128–133.
41. Oldenburg,J., vonBrederlow,B., Fregin,A., Rost,S., Wolz,W., Eberl,W., Eber,S., Lenz,E., Schwaab,R., Brackmann,H.H. *et al.* (2000) Congenital deficiency of vitamin K dependent coagulation factors in two families presents as a genetic defect of the vitamin K-epoxide-reductase-complex. *Thromb. Haemost.*, **84**, 937–941.

42. Pelz,H.-J., Rost,S., Hünerberg,M., Fregin,A., Heiberg,A.-C., Baert,K., MacNicoll,A.D., Prescott,C.V., Walker,A.-S., Oldenburg,J. *et al.* (2005) The genetic basis of resistance to anticoagulants in rodents. *Genetics*, **170**, 1839–1847.
43. Fregin,A., Rost,S., Wolz,W., Krebsova,A., Müller,C.R. and Oldenburg,J. (2002) Homozygosity mapping of a second gene locus for hereditary combined deficiency of vitamin K-dependent clotting factors to the centromeric region of chromosome 16. *Blood*, **100**, 3229–3232.
44. Tie,J.-K., Nicchitta,C., vonHeijne,G. and Stafford,D.W. (2005) Membrane topology mapping of vitamin K epoxide reductase by in vitro translation/cotranslocation. *J. Biol. Chem.*, **280**, 16410–16416.
45. Rost,S., Fregin,A., Ivaskevicius,V., Conzelmann,E., Hürtnagel,K., Pelz,H.-J., Lappegard,K., Seifried,E., Scharrer,I., Tuddenham,E.G.D. *et al.* (2004) Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature*, **427**, 537–541.
46. Li,T., Chang,C.-Y., Jin,D.-Y., Lin,P.-J., Khvorova,A. and Stafford,D.W. (2004) Identification of the gene for vitamin K epoxide reductase. *Nature*, **427**, 541–544.
47. Brown,M., Hughey,R., Krogh,A., Mian,I.S., Sjlander,K. and Haussler,D. (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1**, 47–55.
48. Sjlander,K., Karplus,K., Brown,M., Hughey,R., Krogh,A., Mian,I.S. and Haussler,D. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.
49. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press.
50. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.