BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–8

A visual framework for sequence analysis using n-grams and spectral rearrangement

Stefan R. Maetschke [1,2,4], Karin S. Kassahn [1,2,4] Jasmyn A. Dunn [1], Siew P. Han [3], Eva Z. Curley [1], Katryn J. Stacey [1,3] and Mark A. Ragan [1,2*]

[1]The University of Queensland, Institute for Molecular Bioscience, Brisbane, QLD 4072, Australia
[2]ARC Centre of Excellence in Bioinformatics
[3]The University of Queensland, School of Chemistry and Molecular Biosciences, Brisbane, QLD 4072, Australia
[4]these authors contributed equally

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Protein sequences are often composed of regions that have distinct evolutionary histories as a consequence of domain shuffling, recombination, or gene conversion. New approaches are required to discover, visualize and analyze these sequence regions and thus enable a better understanding of protein evolution.

Results: Here we have developed an alignment-free and visual approach to analyze sequence relationships. We use the number of shared n-grams between sequences as a measure of sequence similarity and rearrange the resulting affinity matrix applying a spectral technique. Heat maps of the affinity matrix are employed to identify and visualize clusters of related sequences or outliers, while n-gram based dot plots and conservation profiles allow detailed analysis of similarities among selected sequences. Using this approach we have identified signatures of domain shuffling in an otherwise poorly characterized family, and homology clusters in another. We conclude that this approach may be generally useful as a framework to analyze related, but highly divergent protein sequences. It is particularly useful as a fast method to study sequence relationships prior to much more time-consuming multiple sequence alignment and phylogenetic analysis.

Availability: A software implementation (MOSAIC) of the framework described here can be downloaded from
`http://bioinformatics.org.au/mosaic/`

Contact: m.ragan@uq.edu.au

1 INTRODUCTION

Eukaryotic proteins often evolve in a modular fashion, acquiring protein domains from related or unrelated sequences via domain shuffling and non-homologous recombination (Schmidt and Davies, 2007; Patthy, 1999; Vogel *et al.*, 2004). In this way, protein families can gain or lose protein domains over time. Domain shuffling is a prominent feature of eukaryotic genomes (Kaessmann *et al.*, 2002), resulting in many sequences having composite evolutionary histories (Vogel *et al.*, 2004). Here, we use the term *reticulate* to refer to these types of sequence relationships which resemble a network rather than a hierarchical, tree-like structure. To date there have been few systematic attempts to identify and characterize reticulate sequence relationships within individual protein families; exceptions include the Markov clusters of homologous subsequences (Wong and Ragan, 2008), the use of intron phase as a signature of domain shuffling (Kaessmann *et al.*, 2002), and comparison of domain combinations across animal taxa (Kawashima *et al.*, 2009). Nonetheless the significance of the modular nature of protein evolution remains largely unexplored. In prokaryotes, lateral genetic transfer can as well create novel recombinant sequences that are composed of subsequences of distinct evolutionary origin (Chan *et al.*, 2009). Identifying reticulate sequence relationships is difficult with established methods.

At present, the predominant approach for studying protein sequence relationships is based on multiple sequence alignment and phylogenetic inference (Baldauf, 2003), but the phylogenetic analysis of large protein families is computationally expensive, and protein families which have been subject to domain shuffling do not lend themselves to a tree-based representation. A number of efforts have thus explored phylogenetic network representations in order to better capture reticulate sequence relationships (Makarenkov, 2001; Bryant and Moulton, 2004; Cardona *et al.*, 2008), but these diagrams are difficult to interpret for larger families and more-complicated reticulate relationships, and the inference of network topologies is even more computationally expensive that that of trees.

In this context, spectral clustering offers an attractive alternative for finding related sequences via clustering (Shi *et al.*, 2000; Ng *et al.*, 2001). Based on the eigenvector decomposition of a similarity matrix, it has already been used successfully to cluster biological sequence data utilizing BLAST similarity scores (Pentney *et al.*, 2005). Furthermore, spectral clustering performed favorably in comparison to three other, single-linkage clustering algorithms when applied to cluster sequences from the SCOP database based on BLAST scores (Pentney *et al.*, 2005). Similarly, Paccanaro *et al.*, 2006 performed spectral clustering on sequences extracted from SCOP to group homologous sequences, but again relied on BLAST alignments and E-values to measure sequence similarity.

*to whom correspondence should be addressed

Alignment-free approaches, in contrast, offer considerable speed advantages and avoid problems associated with sequence rearrangements and the modular makeup of eukaryotic proteins. Alignment-free approaches can be based on comparisons using n-mers, words or patterns, with the former having speed advantages over the latter two (Vinga *et al.*, 2003). Importantly, alignment-free approaches can provide accurate estimates of genetic distance (Höhl *et al.*, 2006; Höhl and Ragan, 2007).

There is thus an opportunity for the development of new approaches to study sequence relationships, especially approaches that can help identify and explain reticulate sequence relationships that reflect the modular nature of protein evolution. Here, we have developed a fast and interactive visual framework for exploring sequence relationships and introduce an alignment-free measure of sequence similarity that can be applied to a broad set of data types, not only sequences but also structural data. The proposed framework can serve as a basis for more-informed and more-accurate phylogenetic analysis.

2 SYSTEM AND METHODS

Our method computes sequence similarities based on the number of shared short subsequences (n-grams) and organizes the resulting similarity scores within an ordered matrix. Employing a spectral technique, the rows and columns of the affinity matrix are rearranged to allow the visual identification of outliers, clusters of related sequences, and reticulate relationships. We furthermore extend the n-gram based approach and present algorithms to generate dot plots and conservation profiles. A software implementation of this framework (MOSAIC) is available at `http://bioinformatics.org.au/mosaic/`.

2.1 Sequence similarity

We define a sequence s as a tuple of N symbols α_i from a finite alphabet L.

$$s = (\alpha_1, \ldots, \alpha_N) \text{ with } \alpha_i \in L, \tag{1}$$

where L may be any alphabet of interest, composed of amino acid, DNA, RNA or secondary structure symbols for instance. To calculate the similarity between sequences we represent sequences by their n-gram sets and compare these sets. An n-gram is herewith defined as a subsequence of length n extracted from a sequence s at position i:

$$\text{n-gram}_i(s) = (\alpha_i, \ldots, \alpha_{i+n-1}), \tag{2}$$

and for each sequence we construct the set S of overlapping but unique n-grams as

$$S(s) = \{\text{n-gram}_i(s) \mid \forall i \in 1 \ldots \text{N-(n-1)}\}. \tag{3}$$

The n-gram similarity σ between two sequences s_1 and s_2 and their corresponding n-gram sets $S_1 = S(s_1)$ and $S_2 = S(s_2)$ is calculated as

$$\sigma(s_1, s_2) = \frac{|S_1 \cap S_2|}{min(|S_1|, |S_2|)}, \tag{4}$$

with σ being the proportion of n-grams shared by both sequences relative to the smaller n-gram set. Note that σ is always in the interval $[0 \ldots 1]$ and that n-gram similarity, in contrast to BLAST scores, is a symmetric similarity measure. Replacing the $min(.)$ operator by the $max(.)$ operator leads to a sequence comparison resembling the differences between a local versus global alignment. Section 6 of the Supplementary Material explains the effect in more detail.

The simplicity of the sequence similarity measure as defined in Equation 4 ensures interpretability of results and fast computation. Specifically, its computation time scales linearly with sequence length. For sequences with a length distribution similar to that of the sequences analyzed here, simulations show that n-grams (on average) are essentially unique for $n > 3$ within amino acid sequences and for $n > 6$ in DNA/RNA sequences (Supplementary Material, Section 5), and thus represent useful units for comparisons of sequence similarity. We found that for amino acid sequences an n-gram size of 4 and for DNA sequence an n-gram size of 12 represent a suitable trade-off between specificity and sensitivity (Supplementary Material, Section 4).

2.2 Spectral rearrangement

Typically, spectral clustering algorithms operate on points in Euclidean space as input, calculate pairwise distances, and perform a spectral decomposition of the affinity matrix derived from point distances. The resulting eigenvectors (or parts thereof) are evaluated by k-means to identify clusters within the input data (Ng *et al.*, 2001; Shi *et al.*, 2000). The MOSAIC framework employs spectral clustering with two significant differences.

First, instead of the Euclidean metric between points, n-gram similarity between sequences is calculated. This, however, requires that n-gram similarities $\sigma(s_i, s_j)$ are transformed into distances $d(s_i, s_j)$, which is readily achieved as follows:

$$d(s_i, s_j) = 1 - \sigma(s_i, s_j). \tag{5}$$

Second, we eliminate the final k-means step of spectral clustering but instead rearrange the rows and columns of the affinity matrix in a way that brings similar sequences together and allows the visual identification of clusters, outliers and other special cases. This *spectral rearrangement* avoids a decision upon the exact number of clusters to detect – a decision which is typically difficult to make *a priori*. Furthermore, in case of reticulate events, fixing the number of clusters does not appropriately describe the underlying sequence relationships.

Following Ng *et al.* (2001) and Shi *et al.* (2000), we construct an affinity matrix A from all pairwise distances $d(s_i, s_j)$, which accentuates the local neighborhood of sequences within the n-gram similarity space:

$$A_{ij} = e^{-d(s_i, s_j)^2 / 2r^2}, \tag{6}$$

with r being the neighborhood radius of a Gaussian kernel that controls how rapidly the affinity A_{ij} between sequences s_i and s_j decreases. Then the Laplacian L of A is computed as:

$$L = B - A, \tag{7}$$

with $B = diag(\sum_k a_i k)$ being a diagonal matrix constructed from the row or column sum of the affinity matrix.

Note that there are alternative ways to construct the Laplacian (von Luxburg, 2007) such as the normalized ($L = I - B^{-1}A$) or the symmetric Laplacian ($L = I - B^{-1/2}AB^{-1/2}$), and the MOSAIC framework allows one to switch between them.

Finally the Laplacian is decomposed into its eigenvectors V and eigenvalues e. The eigenvector v_2, belonging to the 2nd smallest eigenvalue e_2 of the Laplacian, is called the *Fiedler* vector and is indicative of the cluster structure of the matrix (Fiedler, 1975). By reordering the row and columns of D according to the magnitude of its components, related sequences are grouped together.

Figure 1 shows the affinity matrices and the corresponding Fiedler vectors before (left) and after (right) spectral rearrangement for a set of 36 steroid hormone receptor sequences, composed of four families (MR, GR, PR, AR). The relative phylogenetic position of the four steroid receptor types and which is the most-ancestral receptor type is a topic of current research (Hu *et al.*, 2008) and the MOSAIC plot does not resolve this ambiguity, but does correctly cluster sequences that belong to the same family. The Fiedler vector (Figure 1d) allows one to assess objectively how well the four clusters are separated. Larger steps within the Fiedler vector indicate more robust clusters. For instance, sequences belonging to the MR family are clearly less distinguishable from sequences belonging to the GR family than from members of the AR family. Importantly, this comparison could be made

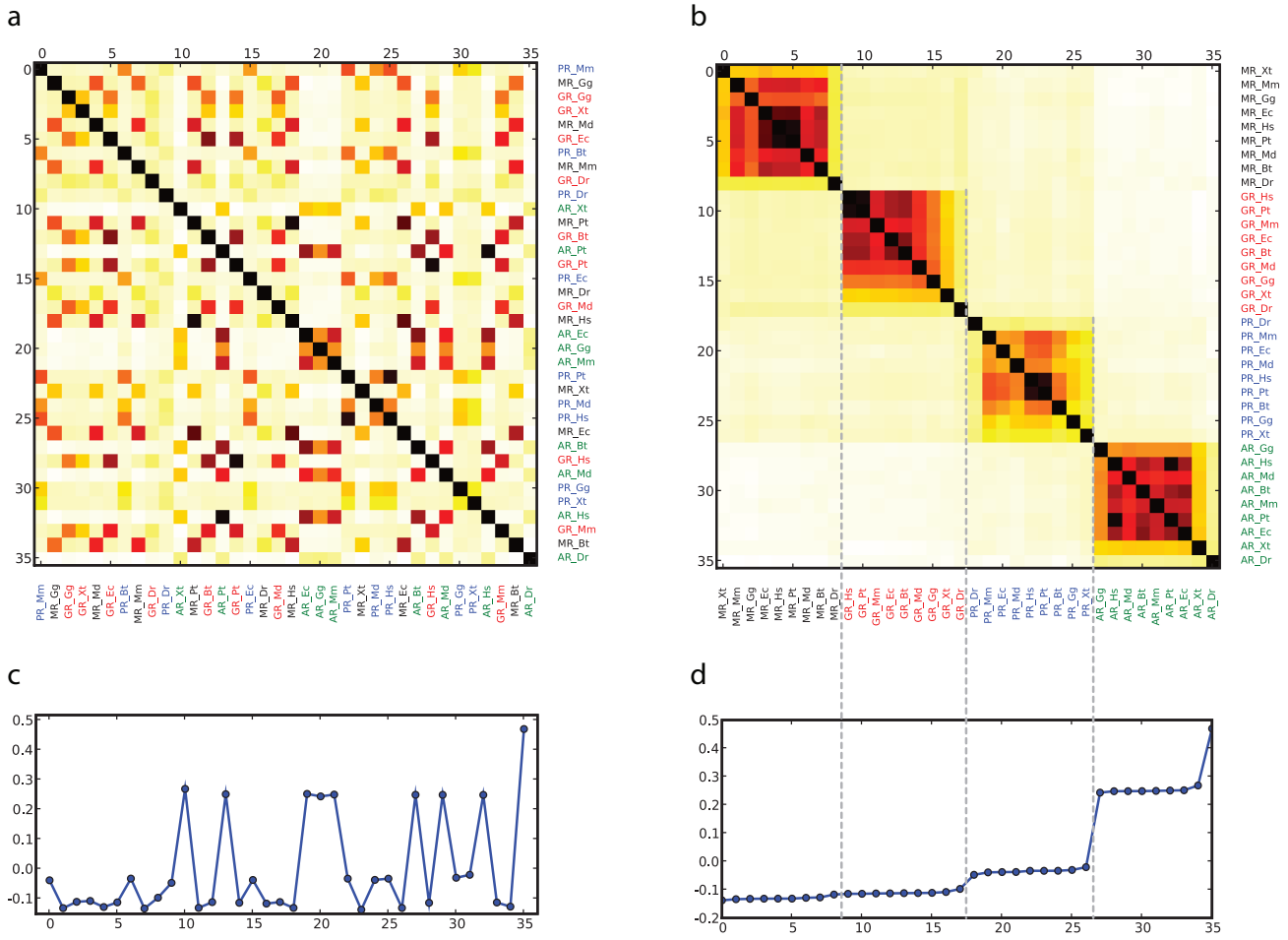

Fig. 1. Spectral rearrangement (non-recursive) of four families of steroid hormone receptors. Top row shows the affinity matrix before (a) and after (b) spectral rearrangement. Bottom row shows the corresponding Fiedler vectors before (c) and after (d) spectral rearrangement. For ease of comparison, sequences belonging to the same receptor family are in the same color. Abbreviations: "AR" androgen receptor, "GR" glucocorticoid receptor, "MR" mineralocorticoid receptor, "PR" progesterone receptor, "Bt" *Bos taurus*, "Dr" *Danio rerio*, "Ec" *Equus caballus*, "Gg" *Gallus gallus*, "Hs' *Homo sapiens*, "Md" *Monodelphis domestica*, "Mm' *Mus musculus*, "Pt" *Pan troglodytes*, "Xt" *Xenopus tropicalis*.

using full-length receptor sequences as opposed to using only selected, conserved domains.

For families with more than two clusters, the clustering is improved considerably by performing a recursive spectral rearrangement. In such a case the affinity matrix is spectrally rearranged as before but then split into two smaller sub-matrices, for which the spectral rearrangement is applied recursively. The split position (row/column index) is derived from the position where the Fiedler vector changes its sign (normalized cut: Shi *et al.*, 2000). The recursion terminates when either the components of the Fiedler vector are all of the same sign (indicating a uniform cluster) or the sub-matrices are reduced to two rows/columns (for further details see Supplementary Material).

2.3 n-gram dot plots

While the overall relationships among multiple sequences can be visualized utilizing the spectral approach outlined above, it is frequently of interest to analyze the similarity between two specific sequences in more detail. To this purpose we extend the traditional concept of a dot plot from individual nucleotides or amino acids to n-grams. Let $M = [m_{i,j}]$ be a matrix that

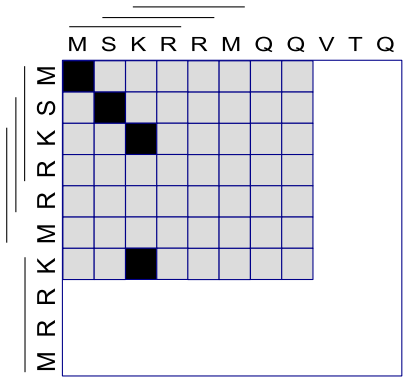

Fig. 2. n-gram dot plot. Matching n-grams (of size four) that occur in both sequences are indicated by horizontal and vertical lines outside the dot plot and by filled squares within the dot plot. Due to an n-gram size of four the last three rows and columns of the plot are empty.

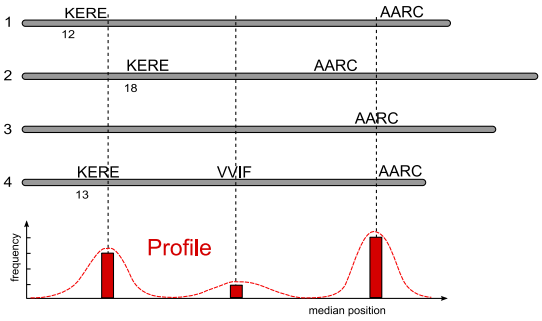

Fig. 3. Illustration of profile generation. Four sequences and some matching 4-grams are shown at the top. "AARC" is conserved at homologous positions in all four sequences, the n-gram "KERE" is conserved in three sequences, and "VVIF" occurs only in one sequences. The profile shown at the bottom is basically the frequency plot of the most frequently occurring n-grams at homologous positions.

represents the dot plot, then a dot will be drawn if the n-grams at positions i and j for two sequences s_1 and s_2 match (Figure 2):

$$m_{i,j} = \begin{cases} 1 \text{ if n-gram}_i(s_1) = \text{n-gram}_j(s_2) \\ 0 \text{ otherwise} \end{cases} \quad (8)$$

$$\forall i, j \in 1 \ldots \text{N-(n-1)}$$

While dot plots for single symbols are typically noisy due to many cross-matches, dot plots for n-grams of reasonable size ($n \geq 3$) are comparatively "clean", since the n-grams of a sequence are largely unique within the sequence. For longer sequences the sparseness of the dot plot has a negative impact on its readability. We therefore construct a compressed matrix representation $M' = [m'_{q,r}]$ of the dot plot that averages over sub-matrices of size $c \times c$ within M:

$$m'_{q,r} = \frac{1}{c^2} \sum_{i=0}^{c-1} \sum_{j=0}^{c-1} m_{q \cdot c - i, r \cdot c - j} \quad (9)$$

$$\forall q, r \in 1 \ldots \text{N/c}$$

Note that the sub-matrices are not overlapping and that the values of M' are not binary anymore, but give a representation of the density of n-gram matches within a sub-matrix. Section 3.3 shows several examples of compressed dot plots.

2.4 n-gram profiles

As a further means to investigate the relationships within a set of sequences visually we compute *n-gram profiles*, which are similar to traditional conservation profiles but combine n-grams and alignment-free methods in a novel way. The method relies on the assumption that n-grams with $n \geq 3$ are generally unique within a single sequence but occur at coherent positions in homologous sequences.

For a given set of sequences, we extract the n-grams that appear in one or more sequences and record their sequence positions (matching n-grams). For instance, n-gram "KERE" may appear at position 12 in sequence 1, at position 18 in sequence 2 and position 13 in sequence 4 but not in sequence 3 (see Figure 3). We then plot the frequency of matching n-grams over their mean positions.

More formally, let C be the set of sequences over which the n-gram profile is to be calculated. We define $\overline{P}(\text{n-gram})$ as the mean position at which the n-gram occurs in all sequences in C. Furthermore, we define $N(\text{n-gram})$ as the number of sequences in C in which the n-gram occurs. The *n-gram profile* is generated by plotting the n-gram frequency $N(\text{n-gram})$ over the mean position $\overline{P}(\text{n-gram})$ at which the n-gram occurs.

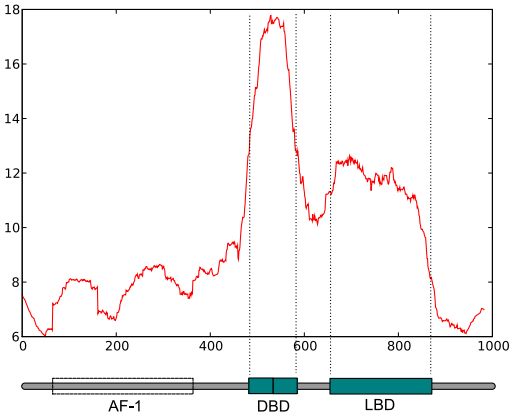

Fig. 4. Profile for steroid hormone receptors. The top graph shows the 4-gram profile derived from 36 steroid hormone receptors sequences. The known domain architecture is illustrated below.

The resulting graph is, however, noisy and two refinements are employed. Firstly, for a specific position $\text{int}(\overline{P})$ we display only $\max(N)$, which is the frequency of the most frequent n-gram at that position, and use the $\text{int}(\cdot)$ function to map a mean position to its nearest integer value. Secondly, we smooth the graph by applying a mean filter.

Figure 4 displays the n-gram profile ($n = 4$) over 36 sequences of steroid hormone receptors (same set as used for Figure 1). Common to steroid hormone receptors are a highly conserved DNA-binding domain (DBD) consisting of two zinc-finger motifs, and a more-variable ligand binding domain (LDB) close to the C-terminus (Whitfield *et al.*, 1999). These two domains can easily be identified within the profile. Considerably less conserved (Lavery *et al.*, 2005) but still visible is the transactivation function domain (AF-1) within the N-terminal section.

3 RESULTS

3.1 Parameter optimization

The approach presented here requires the choice of two parameters, the width r of the Gaussian kernel and the n-gram size n. There is no efficient method to determine r optimally and various rules of thumb are in use. We found that a simple heuristic that sets r to the mean value of the elements of the distance matrix D usually leads to good clustering results.

To optimize the n-gram size n we performed three different types of experiments (Supplementary Material). (1) We calculated the correlation between n-gram sequence similarity and species divergence times for a range of sequence sets, (2) we used a range of protein families of varying relatedness to determine the n-gram size at which the standard deviation of the elements of the distance matrix is maximized, and (3) we determined the area under the ROC curve (AUC) when distinguishing between related and randomly shuffled sequences using n-gram similarity. In all three experiments we found that an n-gram size of 4 was generally a good choice for protein sequences, and that an n-gram size of 12 was usually appropriate for DNA and RNA sequences.

3.2 Application to known cases of gene conversion

To study reticulate protein sequence relationships, we applied our approach to a number of families that are known to have been subject to gene conversion. In these cases, gene conversion was identified on the basis that, at least for part of the gene sequence, gene duplicates within a genome were more similar to each other than to their orthologs in related species (e.g. Kruithof *et al.*, 2007; Perelygin *et al.*, 2008; Annilo *et al.*, 2003).

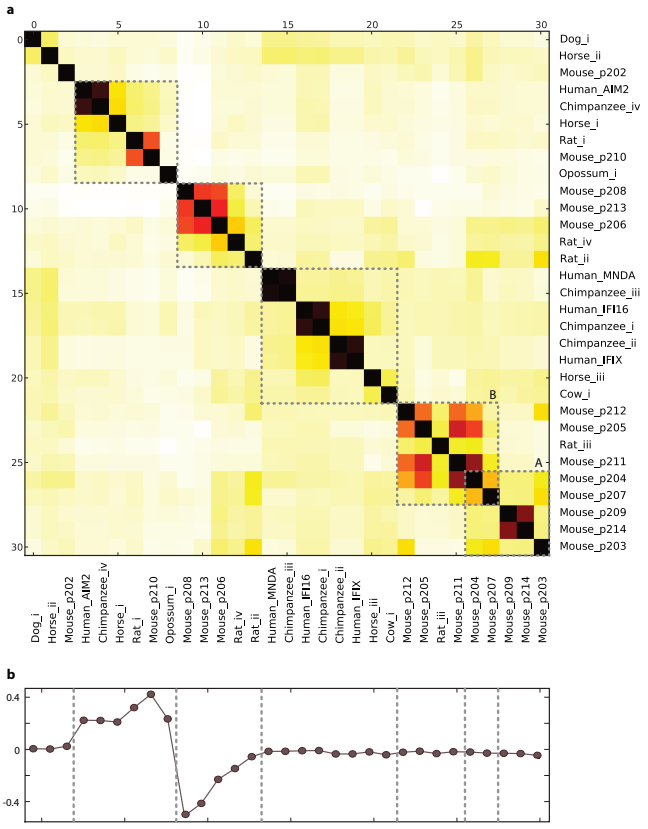

Fig. 5. MOSAIC plot using recursive spectral rearrangement (a) and Fiedler vector (b) of the mammalian HIN-200 protein family. The MOSAIC plot shows many small clusters and an overlap between the weak clusters A and B. HIN-200 protein sequences were taken from NCBI databases and mouse cDNAs for uncharacterised proteins p207, p208, p209, p211, p212, p213 and p214 were cloned and sequenced. Sequences are available upon request.

We used these protein families to test whether our approach was sensitive enough to identify these known reticulate subsequence relationships, and whether n-gram composition and spectral rearrangement provided biologically meaningful results. One of these families included the toll-like receptors TLR1 and TLR6 (Kruithof *et al.*, 2007). Importantly, our approach correctly grouped orthologs, so that mammalian TLR1 and TLR6 sequences formed distinct clusters, while the two mouse receptors TLR1 and TLR6 which had been previously identified as having been subject to gene conversion formed a cluster of their own (Supplementary Figure 1).

Dot plots visualizing pairwise sequence relationships identified a subsequence in TLR1 and TLR6 that shared greater similarity between these mouse paralogs than with their respective orthologs in other mammals, a signature characteristic of gene conversion (Supplementary Figure 1). MOSAIC can thus detect subsequence relationships indicative of gene conversion.

3.3 Application to families that present challenges for alignment-based approaches

The first family is the mammalian HIN-200 protein family, with critical roles in the response to cytoplasmic DNA (Roberts *et al.*, 2009). Phylogenetic analysis of this family is complicated by the presence of highly divergent, "un-alignable" sequence stretches between two conserved protein domains, the pyrin and HIN domains.

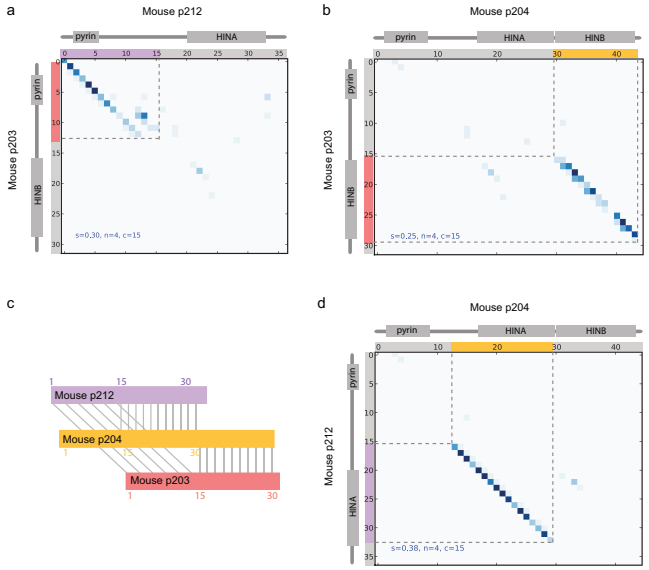

Fig. 6. Dot plots of selected sequence pairs of the HIN protein family. Mouse p212 and p203 only share sequence similarity in the first part of their sequence (a), while p203 is similar to p204 in the second half of its sequence (b). Similarly, p204 is a composite of subsequences that are similar to p203 (b) and p212 (d). The modular makeup of these selected three sequences is illustrated in (c), suggesting that shuffling of subsequences has occurred between family members. The position of recognized Pfam domains is given around each dot plot, demonstrating that the subsequences inferred to have been subject to shuffling extend beyond recognized domain boundaries.

Here, we used MOSAIC to gain insight into the sequence relationships within this family. N-gram composition and spectral rearrangement of sequence similarity indicated the presence of several sub-clusters within this family (Figure 5a). Some of these clusters are well supported, such as the cluster containing the human and mouse Aim2 sequences. Clusters A and B, in contrast, show poor definition (Figure 5a,b). For example, while p203 and p212 belong to clusters A and B, respectively, they both share sequence components with p204 (Figure 6b,d). Pairwise sequence comparisons using dot-plots revealed reticulate sequence relationships. For example, p203 is a composite of subsequences

Fig. 7. Dot plot for sequences p209 and p214. The dot plot shows high n-gram similarity ($\sigma = 0.82$) for an n-gram size of 4 over almost the full length of the two proteins. Note that the dot plot is compressed by a factor of 15 ($c = 15$) to allow better recognition of the main diagonal.

that match p204 and p212, while p212 is a composite of subsequences that match p204 and p203 (Figure 6c), suggesting domain shuffling between family members. Two other sequences in cluster B, p209 and p214, share high sequence similarity across almost the full length of the protein, indicating very recent duplication (Figure 7).

Another set of proteins that poses challenges for phylogenetic analysis includes the hnRNP proteins, a heterogeneous group of nuclear proteins that regulate transcription and splicing (Dreyfuss *et al.*, 1993). While these sequences are often grouped as a family, variation in domain content raises doubts over whether members of this family are truly homologous, and multiple sequence alignment is difficult for these sequences (S.P. Han and Y.H. Tang, unpublished). Using our approach, we identified several clusters of putatively related sequences (Supplementary Figure 2). Beyond these "homology clusters" we found very little similarity between sequences, suggesting that a revision of the so-called hnRNP family is warranted to exclude non-homologous sequences.

4 DISCUSSION

We present here a framework to analyze sequence relationships that consists of three basic elements: n-gram based sequence similarity measurement, spectral decomposition and affinity matrix reordering. These elements in combination yield distinctive advantages for identifying composite evolutionary histories in protein sequences in comparison to traditional methods, such as hierarchical clustering based on BLAST scores, for instance.

4.1 n-gram similarity

The n-gram similarity introduced above defines a symmetric similarity measure – in contrast to BLAST or Smith-Waterman scores –, which is an advantageous property for many applications and specifically allows us to perform a spectral rearrangement of the similarity matrix. The n-gram similarity of two sequences can be calculated in linear time, while traditional BLAST or Smith-Waterman sequence alignments are of quadratic order with respect to sequence length. As an alignment-free method, n-gram similarity reports appropriate similarity scores when classical alignment-based methods fail, e.g. in the case of domain rearrangement. Furthermore, alignment-free methods have been shown to perform as well as alignment-based approaches, even with families that present little challenge to multiple sequence alignment (Höhl and Ragan, 2007). Finally our method makes few assumptions about the processes of evolution and their effects on sequence similarity. It can be applied equally to amino acid, DNA, RNA or other types of sequences.

Other possible approaches, such as computing sequence similarities based on predicted motifs or domains (e.g. via Pfam, InterPro), suffer from limited coverage (approximately 70% for eukaryotic proteomes) and are considerably slower and more-complex than n-gram similarity, but may be of higher specificity.

Note that our method is currently not robust towards point mutations as only perfect n-gram matches are counted to measure sequence similarity. However, n-gram similarity can easily be extended to more sensitive, e.g. by allowing n-grams with mismatches, utilizing n-grams with sub-alphabets or exploiting n-grams of different sizes. Furthermore, n-grams can even be derived from secondary or tertiary structure information to compute similarities between RNA or protein structures.

4.2 n-gram profiles

Traditional sequence alignments such as Smith-Waterman or Needleman-Wunsch utilize dynamic programming to determine columns and/or blocks of matching residues in a set of sequences. From the number of conserved residues within a column, a conservation profile can be derived to identify positions of high or low conservation.

The n-gram based conservation profiles proposed here have the same purpose, but perform n-gram matching in place of dynamic programming to identify regions of conservation. Furthermore, apart from identifying identical n-grams in multiple sequences, no sequence alignment in the traditional sense is required. As a consequence, our algorithm a) has no gap or extension penalties, b) is of linear complexity, whereas dynamic programming is of quadratic complexity with respect to sequence length, and c) can generate useful profiles in situations where traditional alignments fail - for instance, in the case of domain swapping. Note, however, that an n-gram profile with strong conservation, as shown in Figure 4, usually implies that there are alignable (sub)sequences.

The n-gram conservation profiles presented in this paper are also different from entropy or complexity profiles (Vinga *et al.*, 2007; Troyanskaya *et al.*, 2002; Crochemore *et al.*, 1999; Oliver *et al.*, 1993). The latter use a sliding-window approach to determine regions of high or low complexity/entropy within a single (DNA) sequence. These methods are well-suited to search for deviation from randomness and aim to identify functional or regulatory regions, e.g. promoter regions or repetitive regions within a sequence. Conservation profiles, on the other hand, strive to determine regions of sequence conservation across multiple (related) sequences. A set of sequences may exhibit very low complexity/entropy, but still produce very strong (n-gram) conservation profiles.

To summarize, n-gram conservation profiles mesh well with n-gram similarity and n-gram dot plots. They are alignment-free, faster to calculate than traditional conservation profiles,

and better-suited for analysis of complex sequence relationships such as those caused by reticulate events. On the other hand, n-gram conservation profiles (in contrast to profiles based on multiple sequence alignments) do not allow one to determine the conservation of individual residues.

4.3 Spectral rearrangement

To identify clusters of related sequences we followed a spectral approach, which has been shown to be superior to traditional clustering methods in many applications (Ng *et al.*, 2001). Specifically, spectral clustering avoids both the difficulties that k-means clustering encounters in the case of non-spherical clusters, and the chaining problem inherent to single-linkage hierarchical clustering. The optimization problem underlying spectral clustering is unimodal and consequently results in a unique solution of the clustering problem.

Spectral clustering relies on the eigenvector decomposition of the Laplacian. The computational complexity of this operation is of cubic order $O(n^3)$ for dense matrices. However, sparse matrices can easily be constructed by zeroing small similarities, and this can lead to a significant reduction of computational complexity utilizing Lanczos method (Cullum *et al.*, 2002). Furthermore, algorithms have been recently developed to perform an approximative eigenvector decomposition in quasi-linear time (Verma *et al.*, 2001; Sakai *et al.*, 2009).

MOSAIC employs recursive spectral bisection for the matrix rearrangement, which has been shown to generate good partitions in comparison to other methods but in its simplest form is computationally rather expensive (Barnard *et al.*, 1993). Significantly faster implementations such as *multilevel graph partitioning* are available (Barnard *et al.*, 1993; Karypis *et al.*, 1999). Our implementation is comparatively slow but also considerably simpler and sufficiently fast (a few seconds on a standard PC) for the comparatively small data sets that are typically visualized within a MOSAIC plot.

4.4 Other visualizations

We also trialled graph-based methods for visualizing sequence relationships, but found that graphs for larger sets of sequences were unintelligible, creating "hairballs" of difficult-to-decipher relationships (Supplementary Figure 3). In addition, the graph display can differ substantially depending on the layout chosen. The commonly used spring layout, for example, requires the time-consuming optimization of a non-linear problem that has multiple optima. The resulting layout minimizes the squared differences between displayed and true distances, but individual distances are almost always distorted. In contrast, a matrix-rendering approach as employed here is fast, has a single solution and shows true, undistorted distances.

Finally, the software implementation MOSAIC of the method described here enables visualization of subsequence relationships using dot plots and conservation profiles. The dot plots are particularly useful for browsing pair-wise sequence relationships, for example to detect stretches of sequence homology among largely divergent sequences (illustrated here for the hnRNP family) or to explore sequence relationships for signatures of reticulate events (as illustrated for the HIN-200 protein family). The n-gram conservation profiles can be used to identify the position of commonly recurring subsequences even in the case of sequences that are impossible to align using traditional approaches.

4.5 Scalability

In this paper we have successfully applied the MOSAIC framework to data sets with no more than 50 sequences. However, since n-gram similarity can be calculated in linear time, it can be used to analyze much larger sequence sets; we have, for instance, studied complete bacterial proteomes using standard PCs (data not shown). Similarly, efficient algorithms to compute spectral rearrangement are available (see Section 4.3) and we analyzed large subsets of the SCOP database using spectral rearrangement (data not shown).

One limiting factor concerning the evaluation of larger data sets lies with the visual inspection of the MOSAIC plot. Beyond a few hundred sequences the MOSAIC plot loses its advantage of providing both overview and details in a single picture. While it is possible to generate a compressed MOSAIC plot (similar to the compressed dot plot described in Section 2.3), this would render the finer details invisible. The current implementation provides the functionality to zoom into rectangular regions of a (potentially large) MOSAIC plot to study sequence clusters in detail.

5 CONCLUSION

We have presented a framework which can be applied to a broad range of sequence data types and that consistently utilizes n-grams instead of alignment based methods for the visual analysis of sequence relationships, consisting of n-gram based dot plots, conservation profiles and sequence similarity heat maps. We provide the software application MOSAIC, an implementation of this framework, that allows to interactively navigate and study sequence relationships. Although the MOSAIC software makes no quantitative statements, we found it to be very useful for the qualitative analysis of protein families.

Applying our approach to three families that present severe challenges to traditional phylogenetic analysis, we identified signatures of domain shuffling in one of the families and delineated "homology clusters" in the other family, while we were able to recapitulate known cases of gene conversion in the third. While understanding sequence relationships is useful in its own right, it also forms the basis for more-informed phylogenetic analysis. For example, in the presence of domain shuffling, sequences can be split into relevant subsequences before phylogenetic analysis. Similarly, exclusion of outlying or non-homologous sequences will avoid problems with erroneous inference of multiple sequence alignments.

More generally, gene conversion, non-allelic homologous recombination, domain shuffling and other reticulate events are increasingly recognized as having shaped the evolution of many protein families, and there is hence a need to develop new frameworks for analyzing and visualizing the relationships between sequences in such families. The method presented here provides a general framework for exploring sequence relationships and can be easily extended or adjusted for application to specific types of sequence comparisons.

FUNDING

We acknowledge funding from the Australian Research Council Centre of Excellence in Bioinformatics, CE0348221, and National Health and Medical Research Council grant 455920.

REFERENCES

Annilo, T., Chen, Z., Shulenin, S., Dean, M. (2003) Evolutionary analysis of a cluster of ATP-binding cassette (ABC) genes. *Mammalian Genome*, **14**, 7-20.

Baldauf, S.L. (2003) Phylogeny for the faint of heart: a tutorial. *Trends in Genetics*, **19**, 345-351.

Barnard, S.T., Simon, H.D. (1993) A fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. *Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing*, 711-718.

Bryant, D., Moulton, V. (2004) Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**, 255-265.

Cardona, G., Llabres, M., Rossello, F., Valiente, G. (2008) A distance metric for a class of tree-sibling phylogenetic networks. *Bioinformatics*, **24**, 1481-1488.

Chan, C.X., Darling, A.E., Beiko, R.G., Ragan, M.A. (2009) Are protein domains modules of lateral genetic transfer? *PLoS ONE*, **4**, e4524.

Crochemore, M., Vèrin, R. (1999) Zones of Low Entropy in Genomic Sequences. *Computers & Chemistry*, **3-4**, 275-282.

Cullum, J.K, Willoughby, R.A. (2002) Lanczos algorithms for large symmetric eigenvalue computations. *Classics in Applied Mathematics*, **41**, Cambridge University Press.

Dreyfuss, G., Matunis, M.J., Pinol-Roma, S., Burd, C.G. (1993) hnRNP proteins and the biogenesis of mRNA. *Annual Reviews of Biochemistry*, **62**, 289-321.

Fiedler, M. (1975) A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, **25**, 619633.

Höhl, M., Ragan, M.A. (2007) Is multiple sequence alignment required for accurate inference of phylogeny? *Systematic Biology*, **56**, 206-221.

Höhl, M., Rigoutsos, I., Ragan, M.A. (2006) Pattern-based phylogenetic distance estimation and tree reconstruction. *Evolutionary Bioinformatics*, **2**, 357-373.

Hu, X., Funder, J.W. (2008) The evolution of mineralocorticoid receptors. *Molecular Endocrinology*, **20**, 1471-1478.

Kaessmann, H., Zollner, S., Nekrutenko, A., Li, W. (2002) Signatures of domain shuffling in the human genome. *Genome Research*, **12**, 1642-1650.

Karypis, G., Kumar, V. (1999) A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, **20**, 359-392.

Kawashima, T., Kawashima, S., Tanaka, C., Murai, M., Yoneda, M., Putnam, N.H., Rokhsar, D.S., Kanehisa, M., Satoh, N., Wada, H. (2009) Domain shuffling and the evolution of vertebrates. *Genome Research*, **19**, 1393-1403.

Kruithof, E.K.O., Satta, N., Liu, J.W., Dunoyer-Geindre, S., Fish, R.J. (2007) Gene conversion limits divergence of mammalian TLR1 and TLR6. *BMC Evolutionary Biology*, **7**, 148.

Lavery, D.N., McEwan, I.J. (2005) Structure and function of steroid receptor AF1 transactivation domains: induction of active conformations. *Biochemical Journal*, **391**, 449-464.

Makarenkov, V. (2001) T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, **17**, 664-668.

Ng, A.Y., Jordan, M.I., Weiss, Y. (2001) On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems*, **14**, 849-856.

Oliver, J.L., Bernaola-Galván, P., Guerrero-Garciá, J., Román-Roldán, R. (1993) Entropic profiles of DNA sequences through chaos-game-derived images. *Journal of Theoretic Biology*, **160(4)**, 457-470.

Paccanaro, A., Casbon, J.A., Saqi, M.A.S. (2006) Spectral clustering of protein sequences. *Nucleic Acids Research*, **34**, 1571-1580.

Patthy, L. (1999) Protein evolution. Blackwell Science, Oxford.

Pentney, W., Meila, M. (2005) Spectral clustering of biological sequence data. *The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference*, 845-850.

Perelygin, A.A., Zharkikh, A.A., Astakhova, N.M., Lear, T.L., Brinton, M.A. (2008) Concerted evolution of vertebrate CCR2 and CCR5 genes and the origin of a recombinant equine CCR5/2 gene. *Journal of Heredity*, **99**, 500-511.

Posada, D. (2002) Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular Biology and Evolution*, **19**, 708-717.

Roberts, T.L., Idris, A., Dunn, J.A., Kelly, G.M., Burnton, C.M., Hodgson, S., Hardy, L.L., Garceau, V., Sweet, M.J., Ross, I.L., Hume, D.A., Stacey, K.J. (2009) HIN-200 proteins regulate caspase activation in response to foreign cytoplasmic DNA. *Science*, **323**, 1057-1060.

Sakai, T., Imiya, A. (2009) Fast spectral clustering with random projection and sampling. *Lecture Notes in Computer Science*, **5632**, 372-384.

Schmidt, E.E., Davies, C.J. (2007) The origins of polypeptide domains. *Bioessays*, **29**, 262-270.

Shi, J., Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 888-905.

Troyanskaya, O.G., Arbell, O., Koren, Y., Landau, G.M., Bolshoy, A. (2002) Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*, **18(5)**, 679-688.

Verma, D., Meila, M. (2001) A comparison of spectral clustering algorithms. *University of Washington Department of Computer Science Technical Report 03-05-01*.

Vinga, S., Almeida, J. (2003) Alignment-free sequence comparision – a review. *Bioinformatics*, **19**, 513-523.

Vinga, S., Almeida, J.S. (2007) Local Renyi entropic profiles of DNA sequences. *BMC Bioinformatics*, **8**, 393.

Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., Teichmann, S.A. (2004) Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, **14**, 208-216.

von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics and Computing*, **17**, 3954160.

Whitfield, G.K., Jurutka, P.W., Haussler, C.A, Haussler, M.R. (1999) Steroid hormone receptors: Evolution, ligands and molecular basis of biologic function. *Journal of Cellular Biochemistry*, **32-33**, 110-122.

Wiuf, C., Christensen, T., Hein, J. (2001) A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, **18**, 1929-1939.

Wong, S., Ragan, M.A. (2008) MACHOS: Markov clusters of homologous subsequences. *Bioinformatics*, **24**, i77-i85.