MOSAIC: Supplementary material

Stefan R. Maetschke, Karin S. Kassahn, Jasmyn A. Dunn,
Siew P. Han, Eva Z. Curley, Katryn J. Stacey, Mark A. Ragan

January 14, 2010

1 Analysis of toll-like receptors TLR1 and TLR6

To study reticulate sequence relationships, we first applied MOSAIC to a family that has previously been shown to have been subject to gene conversion, the mammalian toll-like receptors TLR1 and TLR6 [Kruithof *et al.*, 2007]. Figure 1 recapitulates the findings of Kruithof *et. al*, showing that the C-terminal regions of mouse TLR1 and TLR6 have stronger sequence similarity to each other than to their orthologs in other mammals. This region encodes most of the intracellular Toll/interleukin-1 receptor (TIR) domain, suggesting that convergent evolution between TLR1 and TLR6 may have been driven by common binding partners that interact with the TIR domains.

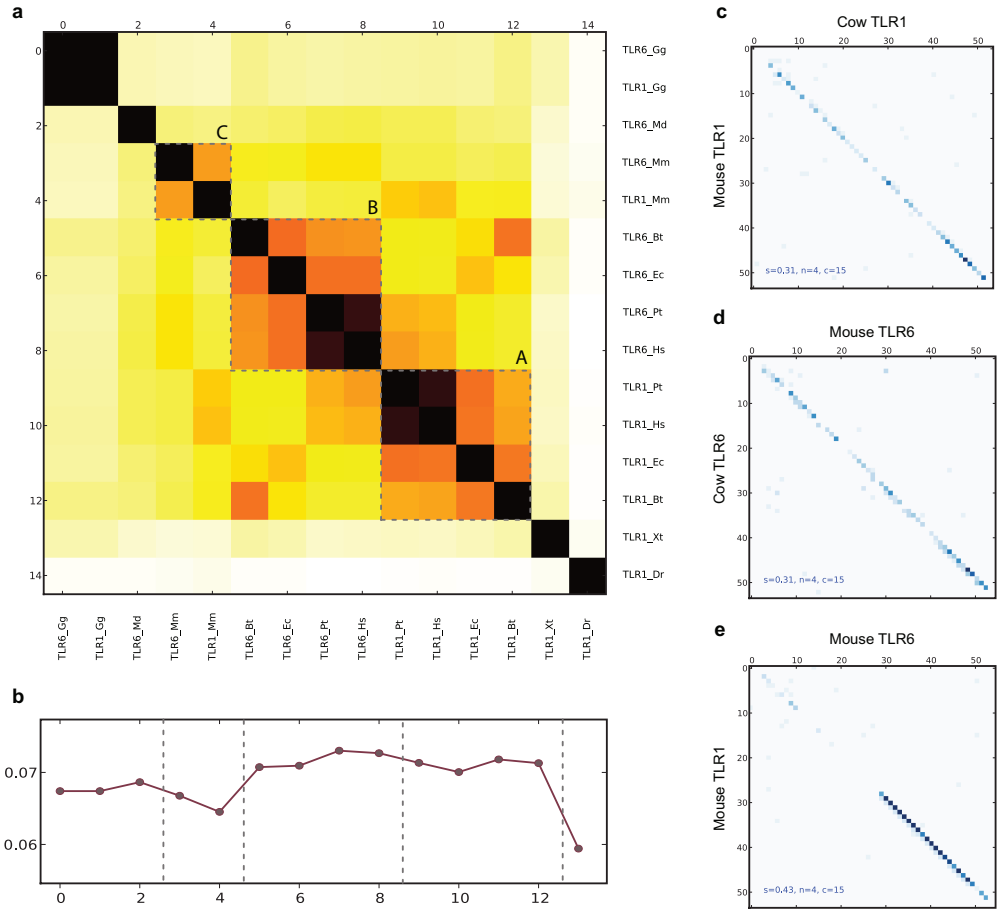

Figure 1: MOSAIC plot (a) and Fiedler vector (b) of toll-like receptor 1 and 6 sequences. Cluster A contains mammalian TLR6 sequences, while cluster B contains TLR1 sequences, but there is some overlap between clusters. Mouse TLR1 and TLR6 sequences (TLR1_Mm and TLR6_Mm) group on their own and separately from the other mammalian TLR sequences (Cluster C), reflecting gene conversion between mouse paralogs. To further investigate the subsequence relationships associated with gene conversion in mouse, mouse TLR1 and TLR6 sequences were compared to their orthologs in cow (c,d) and to each other (e), demonstrating that at least for part of their sequence they share higher sequence similarity than with their mammalian orthologs. Sequences were downloaded from Ensembl (`www.ensembl.org`). Abbreviations: "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*.

2 Analysis of mammalian hnRNP proteins

The mammalian hnRNP proteins have traditionally been grouped as a family, but pose challenges for traditional phylogenetic analysis. Multiple sequence alignment for this group of sequences is impossible due to significant variation in domain content and the presence of repetitive sequence elements, such as glycine-rich regions. Using MOSAIC we delineated clusters of putatively homologous sequences within this heterogeneous group of proteins and identified regions of sequence conservation.

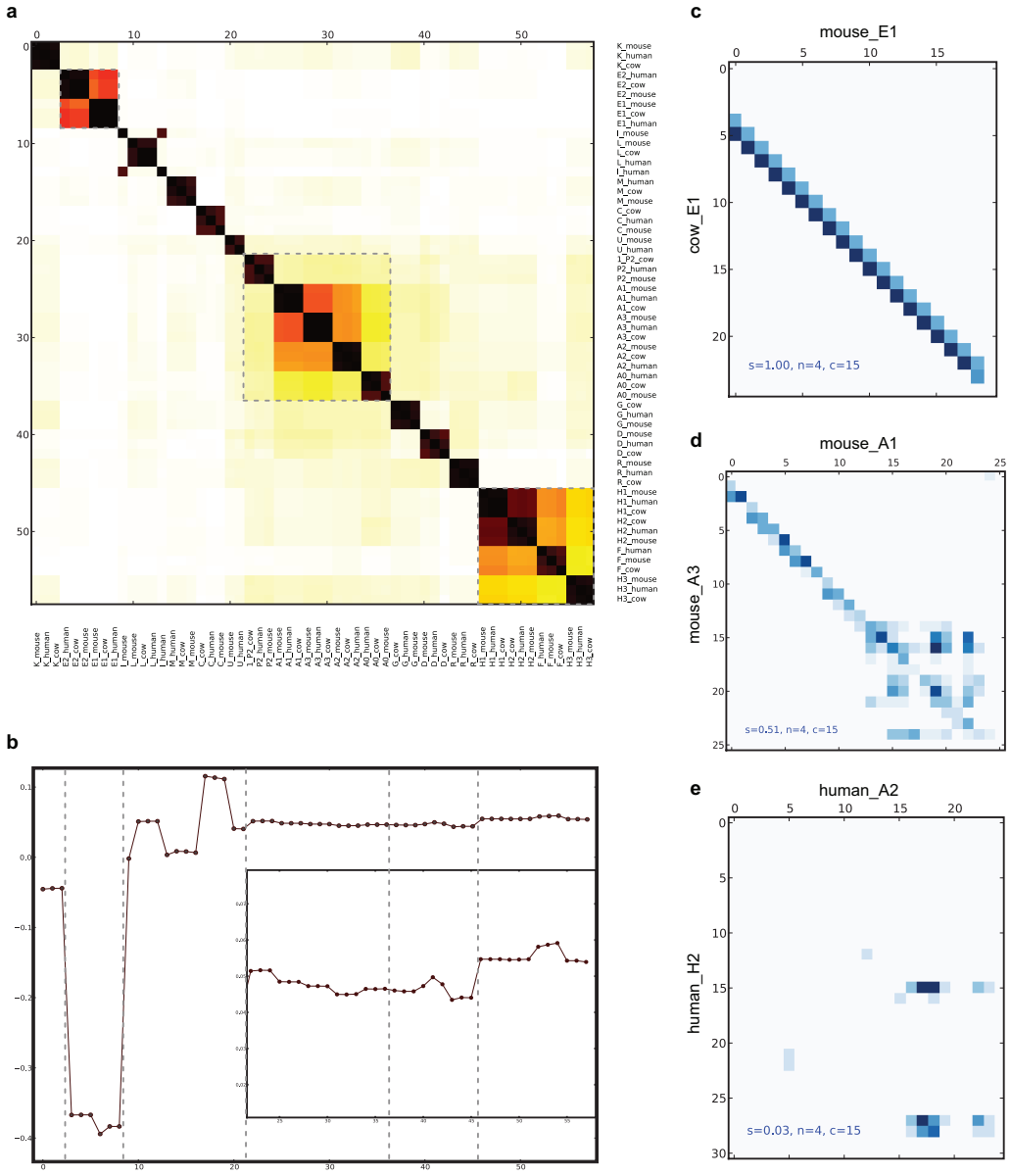

Figure 2: MOSAIC plot (a) and Fiedler vector (b) of mammalian hnRNP proteins. The inset in plot (b) shows a zoomed region of the Fiedler vector. There is little sequence similarity between mammalian hnRNP proteins outside of known orthologs, such as mouse and cow hnRNP E1 (c). Mouse hnRNP A1 and A3 contain a glycine-rich region at the C-terminal of their sequence (d). In contrast, most other so-called hnRNP paralogs share little sequence similarity, as shown for human hnRNP A2 and H2 in (e).

3 Comparison of mosaic plot and graph representations

Reticulate relationships within biological data, such as protein-protein interaction networks, gene regulatory networks and phylogenetic networks, are commonly visualized as graphs, with nodes representing the biological objects and edges describing observed relationships. We found, however, that a graph representation of n-gram similarities between sequences was inferior to a visualization based on the spectral rearrangement of the affinity matrix (MOSAIC plot).

Firstly, identification of clusters of related sequences is usually more difficult from a graph representation than from a MOSAIC plot. Figure 3 displays the same set of sequences (steroid hormone receptors) as a graph with spring layout and as a MOSAIC plot. Note that the colored sequence labels would not be available for sequences where the family association is unknown. The four protein families are easily identified within the MOSAIC plot, but would be very difficult to derive from the graph representation without colored labels.

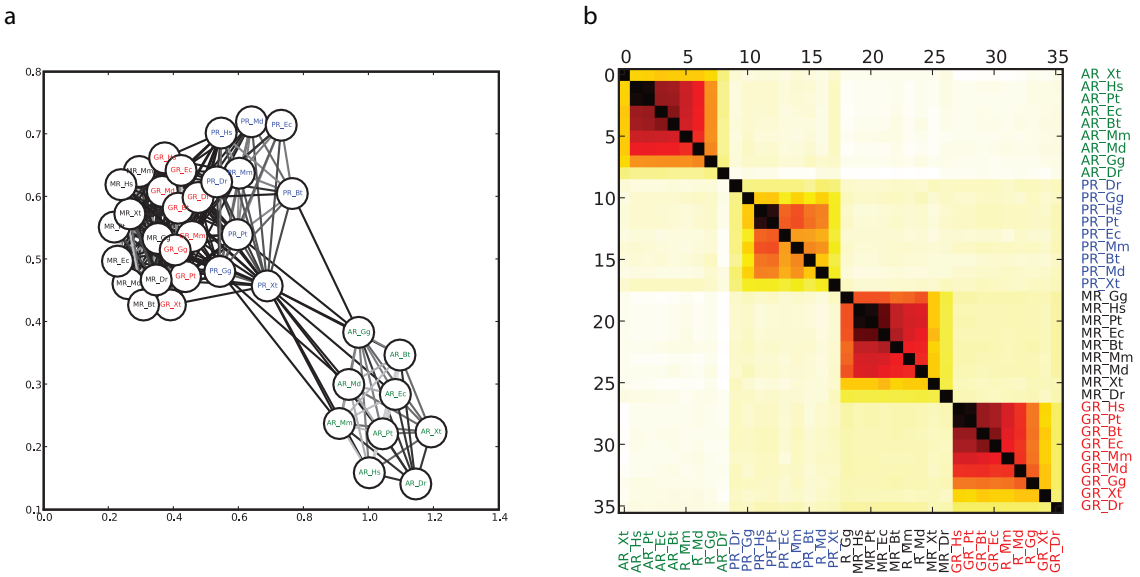

Figure 3: Comparison of a graph-based representation of sequence relationships (a) to the MOSAIC plot using a spectral rearrangement of the affinity matrix (b) for a set of steroid hormone receptor sequences. The four sub-families can easily be identified within the MOSAIC plot.

Secondly, the numerous edges in the graph representation are more confusing than helpful. Their contribution is even more questionable taking into account that the shown distances are distorted for almost all cases and can not be compared. The commonly employed spring layout, which was chosen here to arrange the nodes of the graph, optimizes a quadratic error function derived from the squared differences between the true distances (sequence dissimilarity) and the edge lengths. In contrast, the affinity matrix of the MOSAIC plot is not distorted in this way and distances between sequence pairs (represented as color squares) are comparable.

Thirdly, the optimization of a graph layout typically is a non-linear problem with many local minima. Consequently, each time the optimization is performed a different layout may be calculated. The spectral rearrangement, on the other hand, always yields a unique solution.

4 Optimization of n-gram size

The MOSAIC framework calculates the similarity of sequences using the number of shared n-grams. The discriminative power to distinguish between related and unrelated sequences therefore depends on the choice of an appropriate n-gram size. For small n, sequences become more similar than for large n and there is a trade-off between sensitivity and specificity. In this section we describe three different methods to determine an optimal n-gram size.

4.1 Description of datasets

To optimize the n-gram size, we utilized six sequence datasets representing various levels of sequence divergence and whose phylogenetic relationships have been previously published:

- ATP-binding cassette genes ABCA5, ABCA6, ABCA8, ABCA9, and ABCA10 [Annilo *et al.*, 2003].
- Toll-like receptors TLR1 and TLR6 [Kruithof *et al.*, 2007].
- Actin genes ACTB, ACTR1A and ACTR2 [Goodson and Hawse, 2002].
- Chemokine receptors CCR2, CCR5, and CCR3 [Perelygin *et al.*, 2008].
- HPRT1 and PRTFDC1 gene families [Keebaugh *et al.*, 2007].
- Mineralocorticoid (NR3C2 or MR), glucocorticoid (NR3C1 or GR), androgen (AR), and progesterone (PGR or PR) receptors [Hu and Funder, 2008].

Sequences belonging to these families were downloaded from Ensembl, UniProt or GenBank. Each dataset contained between 17 and 39 sequences and included orthologs across vertebrate species as well as paralogs from within the same species. These sequence datasets included highly-conserved as well as more-variable ones in order to determine the effects of varying rates of evolution on optimal n-gram size.

4.2 Standard deviation of similarity matrix

We first optimized the n-gram size by maximizing the standard deviation of the similarity matrix. The reasoning is that an "interesting" similarity matrix, or a matrix that provides the most information, is one that varies maximally in its elements and allows the greatest distinction between similar and dissimilar sequences.

Figure 4 plots the standard deviation of the similarity matrix over n-gram size for the six different sequence sets described above. Some of these sets, such as Bactin_ARP1_ARP2, contain very distinct families while for others, such as ABC transporters (ABCA), the differences in sequence similarity for different families are less pronounced. These difference in family distinctiveness are reflected by the maximum standard deviation achieved and the corresponding best n. As expected, sequence sets with more-distinct family structures yield higher standard deviations based on slightly longer n-grams. In general, however, the highest standard deviation is reached for an n-gram size of 4.

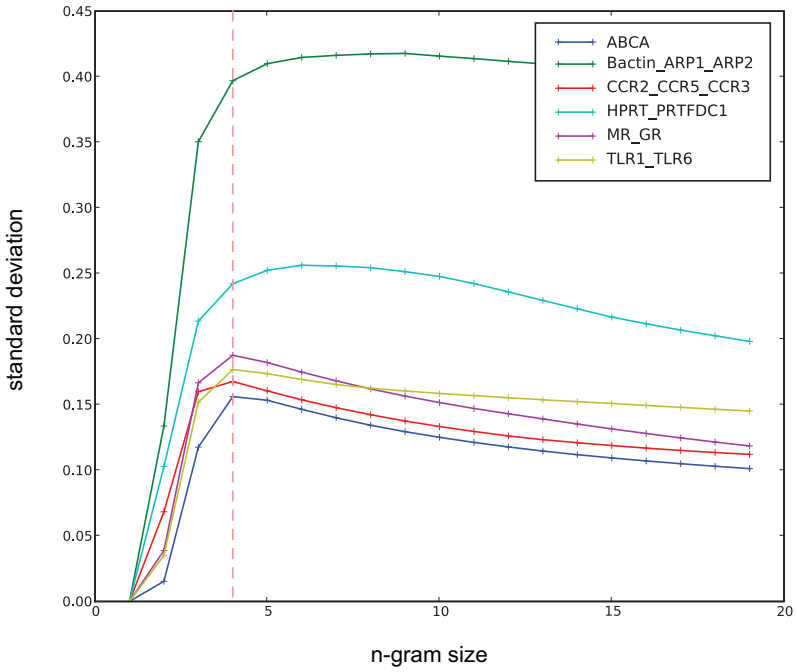

Figure 4: Standard deviation for different sizes of n and six different data sets. The standard deviation typically peaks for an n-gram size of $n = 4$.

4.3 Discrimination between related and unrelated sequences

Another possible approach to optimize the n-gram size is to maximize the power to discriminate between related and unrelated sequences. For this purpose, a set of related sequences, e.g. originating from a protein family, is taken and the pairwise n-gram similarities of all sequences within the set are calculated (first set). A second set of unrelated sequences is generated by randomly shuffling the residues of all sequences within the first set. Then the n-gram similarities between all sequences in the first and the second set are computed. In the case of a perfect similarity metric, all similarities measured within the set of related sequences (first set) should be higher than those computed between related (first set) and unrelated sequences (second set).

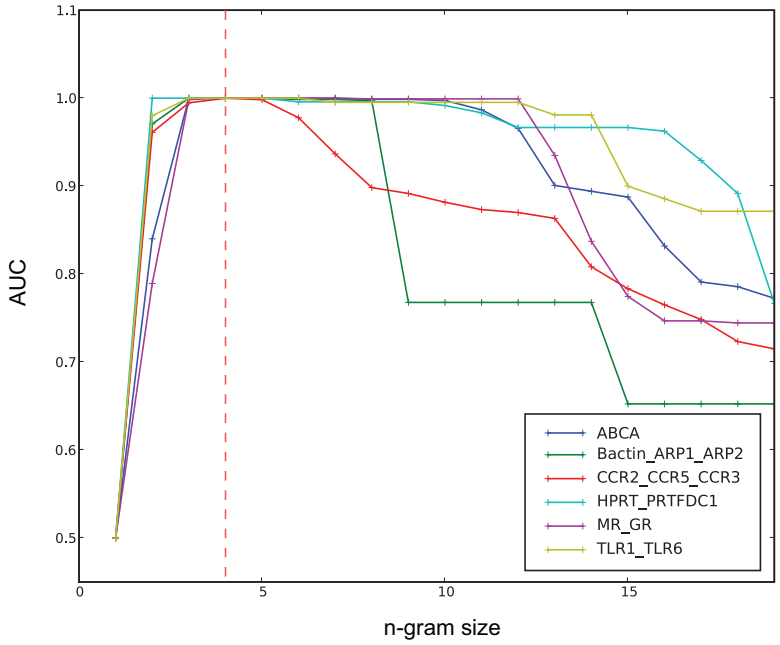

Figure 5: AUC for different sizes of n for six data sets. The highest AUC is typically achieved for $n = 4$.

Based on these similarity scores the AUC (area under the ROC curve) can be calculated for a given n-gram size. We measured the discriminative power (AUC) over different n-grams sizes for the six data sets described in Section 4.1. Figure 4 shows the corresponding curves, which indicate that an n-gram size of 4 generally leads to the greatest discrimination between related and unrelated sequences.

4.4 Species divergence times

Finally, we evaluated the correlation between n-gram sequence similarity and species divergence times as a means to optimize the n-gram size. Assuming a molecular clock, there should be a linear relationship between genetic distance (estimated via n-gram similarity) and species divergence time. We thus calculated the correlation coefficient between n-gram similarity of sequences and their corresponding divergence times (taken from [Ponting, 2008], see Figure 6).

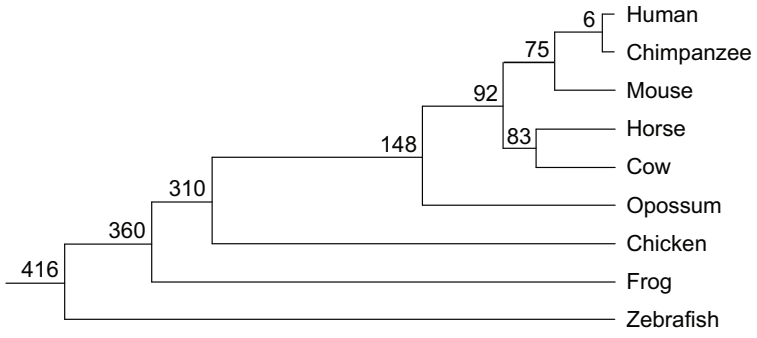

Figure 6: Phylogenetic relationships among nine vertebrate species and divergence times in million years. Divergence times were taken from [Ponting, 2008]. Branch lengths are not to scale.

The correlation coefficients were computed for different n-gram sizes over all sequence subfamilies within the six datasets described in Section 4.1. For instance, Figure 7 shows the correlation between n-gram similarity and species divergence time for the four steroid receptor families, varying the n-gram size between 2 and 19.

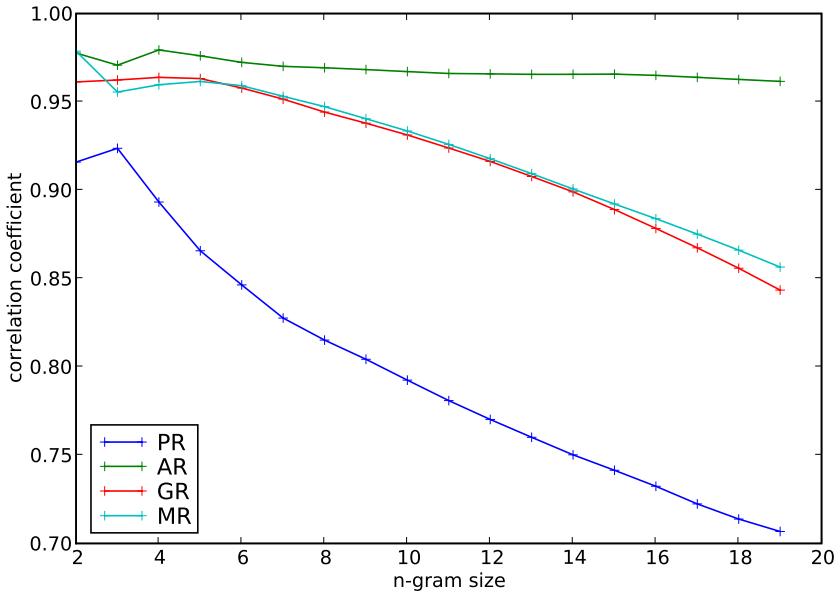

Figure 7: Correlation between species divergence times and n-gram similarity for different n-gram sizes, calculated for sequences from four steroid receptor families.

For other sequence families the corresponding correlation curves can appear quite different but we found that n-gram sizes between 3 and 5 generally yielded the greatest correlation coefficients.

5 Frequency of n-grams

The function of the MOSAIC framework relies on the assumption that n-grams of reasonable size ($n > 3$ for protein sequences) are essentially unique within a sequence. To test this assumption we calculated the frequencies of n-grams within a sequence for different n-gram sizes over the six sequence sets described in Section 4.1.

Figure 8 shows the average frequency of n-grams within a sequence, averaged over all sequences within a set. For n-gram sizes greater than 3 the average frequency closely approximates 1, validating the assumption that n-grams longer than 3 are effectively unique within a sequence. This observation supports our approach to measure the similarity of sequences based on the number of shared n-grams and ignoring the n-gram frequency.

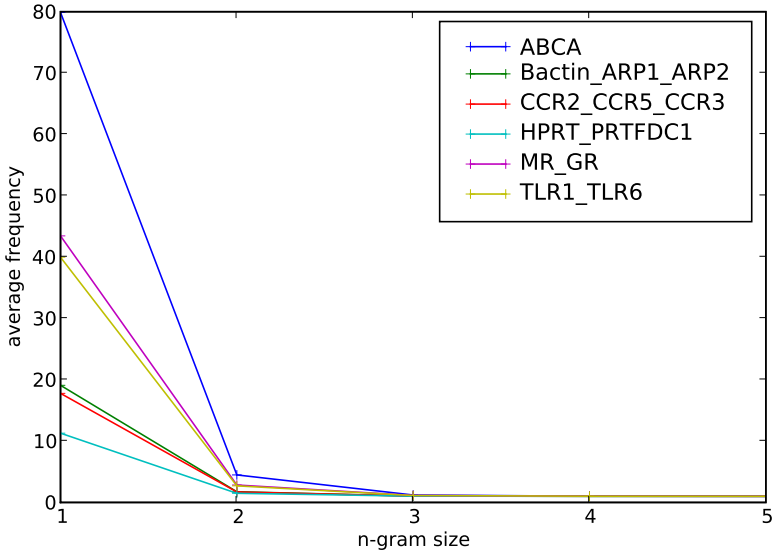

Figure 8: Averaged frequency of n-grams within a sequence over n-gram size for six different sequence sets.

An interesting question to ask is whether the uniqueness of n-grams is specific for the amino acid sequences of our data sets or a salient feature of any amino acid sequence or even of any sequence derived from an amino acid or DNA/RNA alphabet. Table 1 lists the average n-gram frequencies for different n-gram sizes for four different sequence sets.

The *original* set contains all sequences of the six data sets described in Section 4.1. For each of the sequences the average n-gram frequency was calculated and then averaged over the number of sequences. The *randomized* set was created by randomly shuffling the amino acids of the *original* set. The lengths of the sequences and their amino acid distributions were conserved. The *artificial-AA* set was generated by randomly picking amino acids from a uniform distribution but preserving the lengths of the original sequences. Similarly, an artificial DNA sequence set (*artificial-DNA*) was generated by randomly picking nucleotides from a uniform distribution. In this case the sequence lengths of the original sequences were multiplied by a factor of three – taking the codon size into account – to ensure that the frequencies are comparable.

n	1	2	3	4	5	6	7	8	9
original	35.248	2.497	1.099	1.011	1.002	1.001	1.001	1.000	1.000
randomized	35.248	2.471	1.088	1.005	1.000	1.000	1.000	1.000	1.000
artificial-AA	35.228	2.197	1.044	1.002	1.000	1.000	1.000	1.000	1.000
artificial-DNA	528.413	132.041	32.995	8.302	2.453	1.288	1.066	1.016	1.004

Table 1: Average n-gram frequencies for $n = 4$ over the original sequence set, a set of randomized sequences, and two artificial sets composed of amino acid and DNA sequences.

The results in Table 1 show very little differences in the averaged n-gram frequencies for the three amino acid sequence sets, which strongly indicates that the uniqueness of n-grams of size 4 is essentially an inherent feature of any sequence derived from a 20 letter alphabet - assuming a letter distribution not dramatically different from the uniform distribution. Note, however, that the averaged n-gram frequencies of the original data set are generally slightly higher than those of the randomized data set, which in turn are slightly higher than those of the artificial data set. This can be explained by the fact that real biological sequences are more constraint in their composition than randomly shuffled sequences or artificial sequences.

Due to the reduced alphabet size and increased sequence length the averaged n-gram frequencies for the artificial DNA sequence set are considerably higher than the corresponding frequencies for the amino acid sequence sets. However, even in this case we find that for $n > 7$ the average frequency of n-grams finally approximates 1.

The averaged n-gram frequency is a function of three parameters: n-gram size, amino acid distribution, and sequence length. Having established that for n-grams of size 4 and sequences with 'average' amino acid distributions (datasets from Section 4.1) the averaged n-gram frequency is effectively 1, we studied the relationship between sequence length and n-gram frequency. Figure 9 plots the averaged frequency of 4-grams over artificial amino acid sequences of increasing length. It is apparent that even for very long sequences (up to 10,000 amino acids in length) the averaged n-gram frequency can be assumed to be approximately 1.

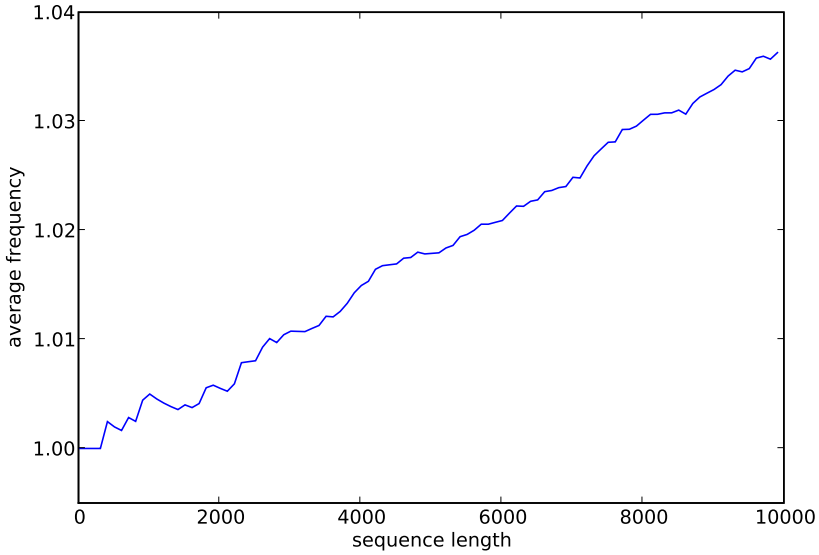

Figure 9: Averaged frequency of 4-grams over artificial amino acid sequences (uniform distribution) of increasing length.

Only on a genome-wide scale, averaged n-gram frequencies climb beyond 1. Figure 10 shows the averaged frequency of 4-grams over artificial amino acid sequences with a length of up to 1

million amino acids, for which the averaged n-gram frequency is approximately 6.

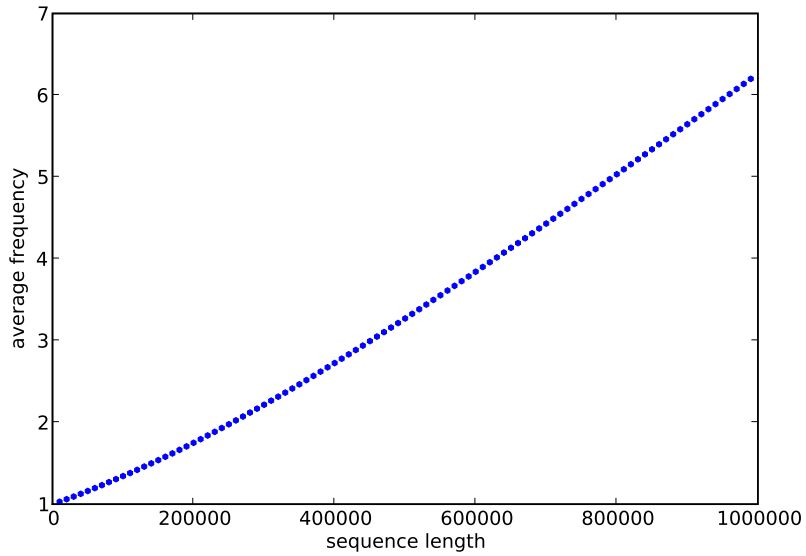

Figure 10: Averaged frequency of 4-grams over artificial amino acid sequences up to 1 million amino acids in length.

6 Minimum versus maximum n-gram similarity

One of the attractive features of the proposed n-gram similarity is that its scores can resemble a local or a global alignment depending on whether the $min(.)$ or the $max(.)$ operator is used. This section explains in more detail the properties of the minimum and the maximum n-gram similarity (Equations 1, 2).

A traditional local alignment method identifies regions of high similarity within long sequences and essentially ignores the non-matching sections of two sequences. A pairwise local alignment of two sequences of varying sequence lengths will therefore return the same score as an alignment of two sequences with identical sequence length. In contrast, a global alignment forces an alignment over the entire length of the paired sequences and differences in the lengths of the aligned sequences affect the alignment score due to gap penalties.

While not an alignment method, the n-gram similarity displays a similar behavior with respect to the impact of varying sequence lengths on similarity scores. When the $min(.)$ operator is used, length differences of the matched sequences are essentially ignored, while use of the $max(.)$ operator takes length differences into account. As an example let us assume two sequences A and B, with sequence A containing a section that perfectly matches a section of a sequence B but none of the n-grams in A match anywhere else in B. Furthermore, let sequence B be twice as long as A. Figure 11 shows a schematic of this example, where sequence A perfectly matches the first half of sequence B. Note however, that this is a simplified example and that the matching n-grams (or blocks of n-grams) do not have to be consecutive.

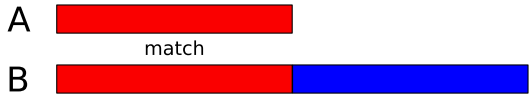

Figure 11: Example of two sequences A and B, with A perfectly matching half of sequence B, and B being twice as long as sequence A.

Ignoring n-gram mismatches at the section borders, and assuming an n-gram frequency of 1 for all n-grams, the minimum n-gram similarity (see Equation 1) for these two sequences would be 1.0. The number of shared n-grams $|A \cap B|$ is equal to the number of n-grams $|A|$ and the minimum $min(|A|, |B|)$ of the numbers of n-grams of A and B is the number of n-grams of $|A|$, resulting in $|A|/|A| = 1$

$$\sigma(A, B)_{min} = \frac{|A \cap B|}{min(|A|, |B|)}, \tag{1}$$

In contrast, using the maximum n-gram similarity (see Equation 2), the denominator changes to $max(|A|, |B|) = |B|$, resulting in $|A|/|B| = |A|/(2 * |A|) = 0.5$

$$\sigma(s_1, s_2)_{max} = \frac{|S_1 \cap S_2|}{max(|S_1|, |S_2|)}, \tag{2}$$

To summarize, the minimum n-gram similarity neglects differences in sequence lengths when calculating similarity scores and in this regard resembles a local alignment approach. In contrast, the maximum n-gram similarity takes length differences into account and thus resembles a global alignment approach.

7 Recursive spectral rearrangement

A spectral rearrangement of a matrix can be achieved by performing an eigenvector decomposition and ordering the rows and columns of the matrix according to the element sizes of the eigenvector belonging to the second smallest eigenvector (Fiedler vector). The rearrangement operation brings similar rows and columns together and allows the identification of clusters. While spectral rearrangement is usually successful in revealing the main clusters within the data, the ordering of elements within a cluster is frequently not optimal.

The intra-cluster organization can be improved by performing a spectral rearrangement for each of the clusters, which can be generalized to a recursive spectral rearrangement. Figure 12 shows the affinity matrices with non-recursive and recursive spectral rearrangement for a set of steroid hormone receptors sequences. While in both cases the four families (MR, GR, AR, PR) are clustered correctly, the ordering of sequences within the clusters is clearly better for the recursive spectral arrangement (compare Figure 12a and Figure 12b).

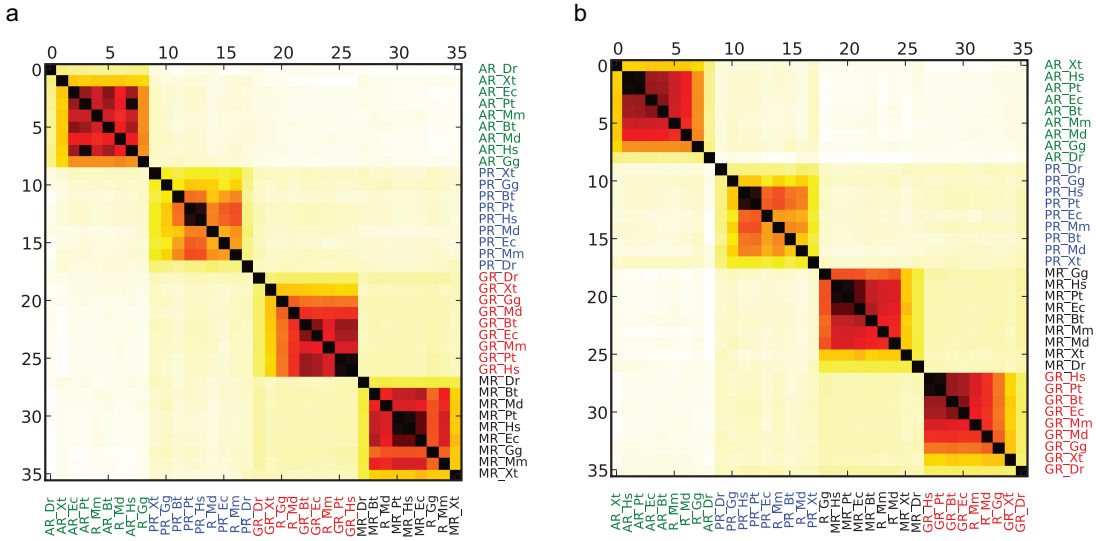

Figure 12: Non-recursive (a) and recursive (b) spectral rearrangement of the affinity matrix for four families of steroid hormone receptors sequences.

Recursive spectral rearrangement is implemented by performing a spectral rearrangement first and then splitting the resulting reordered affinity matrix 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. The recursion terminates when either the components of the Fiedler vector are all of the same sign (indicating a uniform cluster), or there is no further change in the arrangement, or the sub-matrices are reduced to two rows/columns. The following Python code provides the details of the implementation.

```
import numpy

def min_cut(ev):
    """find index where Fiedler vector changes sign
       ev -- Fiedler vector
    """
    for i in xrange(1,len(ev)):
        if ev[i-1] < 0 < ev[i]: return i
    return None   # can't cut: only one cluster


def laplacian(A):
    """calculates the standard laplacian
       A -- affinity matrix
    """
    A[A<0.01] = 0                       # Remove small similarities
    I = identity(len(A))                # Identity matrix
    s = A.sum(axis=0)                   # row sum
    return diag(s) - A


def rearrange(A, o):
    """recursive spectral rearrangement
       A -- affinity matrix
       o -- ordering: list of col/row indices
    """
    if len(o) < 2: return o             # matrix too small, stop recursion
    L = laplacian(A)                    # calc. Laplacian
    w,v = eigh(L)                       # eigenvector decomposition
    ev = v[:,argsort(w)[1]]             # second smallest eigenvector
    fi = argsort(ev)                    # Fiedler vector, sorted
    if list(fi)==range(len(fi)):        # no change in order, stop recursion
        return o                        # return current ordering
    A = A[:,fi][fi,:]                   # matrix rearrangement
    o = o[fi]                           # col and row ordering
    mc = min_cut(ev[fi])                # find min cut
    if not mc: return o                 # uniform Fiedler vector, stop recursion
    o1 = rearrange(A[:mc,:mc], o[:mc]) # rearrange first cluster
    o2 = rearrange(A[mc:,mc:], o[mc:]) # rearrange second cluster
    return concatenate((o1,o2))         # new row/col ordering
```

References

[Annilo *et al.*, 2003] 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.

[Goodson and Hawse, 2002] Goodson, H.V., Hawse, W.F. (2002). Molecular evolution of the actin family. *Journal of Cell Science*, **115**, 2619-2622.

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

[Keebaugh *et al.*, 2007] Keebaugh, A.C., Sullivan, R.T., NISC Comparative Sequencing Program, Thomas, J.W. (2007). Gene duplication and inactivation in the HPRT gene family. *Genomics*, **89**, 134-142.

[Kruithof *et al.*, 2007] 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.

[Perelygin *et al.*, 2008] 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.

[Ponting, 2008] Ponting, C.P. (2008) The functional repertoires of metazoan genomes. *Nature Reviews Genetics*, **9**, 689-698.