The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 189–195

Local Similarities and Clustering of Biological Sequences: New Insights from N-local Decoding

Eduardo Corel[2] Ramzi El Fegalhi[1] Fanny Gérardin[2]
Mark Hoebeke[2] Marc Nadal[1] Alexandre Grossmann[2]
Claudine Devauchelle[2]

[1] Institut de Génétique et Microbiologie, UMR8621 CNRS, Université Paris Sud 11,
Centre Universitaire d'Orsay, bât 409, 91045 Orsay Cedex, France
ramzi.elfeghali,marc.nadal@igmors.u-psud.fr
[2] Laboratoire Statistique et Génome, CNRS UMR 8071, INRA 1152,
Université d'Evry, Tour Evry2, Place des Terrasses, 91034 Evry Cedex, France
corel, grossmann, devauchelle@genopole.cnrs.fr

Introduction

The search for local similarities in sequences is a classical problem in biology, and several methods have been developed for this goal. We herein investigate the N-local decoding method due to Gilles Didier ([2]), in order to classify sequences according to the local similarity segments of a fixed length N that they share. The sites of our sequences are originally occupied by a nucleotide or an amino-acid, and, after the N-local decoding has been applied, these sites are occupied by new symbols (that we call *GD-classes*), which classify the sites according to the composition of their environment in words of length N. This method has already been successfully used to construct trees for the subtyping of HIV/SIV variants ([3]). After recalling the definitions and the original method of the N-local decoding, we will present new developments which aim, on the one hand to allow to exploit the information generated by the decoding, and on the other hand, to tackle the influence of the free parameter N.

1 Methods

In this paper we consider sequences on a finite alphabet $\mathcal{A}$. The general problem of the N-local decoding is concerned with occurrences of identical subwords of fixed length N in sequences.

1.1 Definitions

1.1.1 Sequences et sites

Let $|s|$ be the length of a sequence s, let s_i be its elements for $1 \leq i \leq |s|$, and let $s_{[i,j]}$ denote the subsequence $s_i s_{i+1} \cdots s_j$. A sequence s' is *contained in s at position i* if $s' = s_{[i,i+|s'|-1]}$. Let us consider a set S of sequences over the alphabet $\mathcal{A}$.

A *site* denotes a position in a sequence, in order to distinguish it from the letter it carries. Formally the j-th site of the sequence s is the couple $\sigma = (s, j)$, for $1 \leq j \leq |s|$. If $1 \leq j+k \leq |s|$, we will put $\sigma + k = (s, j+k)$. Let Σ be the set of sites of S.

Let $R \subset \Sigma \times \Sigma$ be a binary relation over Σ, and $\Delta = \{(\sigma, \sigma) \,|\, \sigma \in \Sigma\}$ the *diagonal of* Σ. The transitive closure $\overline{R}$ of $R \cup \Delta$ is the *equivalence relation spanned by R*.

1.1.2 Neighbourhoods

The *N-neighbourhood of site* $\sigma = (s, j)$ is the set of sites $(s, m_N), \ldots, (s, M_N)$ where

- $m_N = \sup(1, j - N + 1)$
- $M_N = \inf(|s|, j + N - 1)$.

It is equivalently the window of width $2N-1$ centered at σ, and possibly truncated at the ends of s. At any rate, the N-neighbourhood of a site is a subsequence of s of length k satisfying $N \leqslant k \leqslant 2N-1$. A word W of length N (hereafter called an *N-word*) is said to be *in relative position ℓ with respect to the site* $\sigma = (s, i)$ if the subsequence $s_{[i-\ell, i-\ell+N-1]}$ coincides with the word W.

1.2 Decodings

Generally speaking a *decoding* of a set S of sequences is a map $f : \Sigma \longrightarrow E$ from the set of sites of S into a finite set E of *states*. If there is no natural way to annotate the states, a decoding is equivalent to a partition P of Σ.

In this section, we recall Gilles Didier's procedure of N-local decoding ([2]). Let us define the *direct similarity relation of order N among sites* $\simeq_N$ as $\sigma \simeq_N \sigma'$ if and only if there exists an N-word W at the same relative position in the N-neighbourhoods of σ and σ'. The transitive closure $\sim_N$ of $\simeq_N$ is an equivalence relation among sites of S. Let $[\sigma]_N$ denote the class modulo $\sim_N$ of a site σ, and P_N the partition of Σ induced by $\sim_N$. This partition is the *N-local decoding of S*.

For $0 \leqslant i \leqslant N-1$ let us define $\overset{i}{\equiv}_N$ the *i-th similarity relation of order N between sites* by $\sigma \overset{i}{\equiv}_N \sigma'$ if and only if σ and σ' are both located at the i-th position of the same N-word W in their respective neighbourhoods. The relation $\sigma \overset{i}{\equiv}_N \sigma'$ is an equivalence relation, for all i, and, obviously, $\sigma \overset{i}{\equiv}_N \sigma'$ holds if and only if $\sigma - i \overset{0}{\equiv}_N \sigma' - i$ does. Therefore, the following lemma is self-evident.

Lemma 1.
For all $N \geqslant 0$, we have

$$\sim_N = \overline{\bigcup_{i=0}^{N-1} \overset{i}{\equiv}_N}.$$

Let S denote now the concatenation of sequences of S obtained by inserting between them different symbols not belonging to the alphabet $\mathcal{A}$, and let $STree(S)$ be the attached suffix tree, whose leaves are thus indexed by the sites of S (and the external symbols). A node of depth N is a common ancestor of two leaves σ and σ' if they are the starting site of a common word of length N, *i. e.* if $\sigma \overset{0}{\equiv}_N \sigma'$.

It is this observation which allows us to rely on the suffix tree construction of Ukkonen ([7]) for computing these equivalence classes, which yields a very fast procedure for the computation of the N-local decoding. For more details, we refer the reader to [2].

We will simply list here a performance array, where we vary the size of the input, the integer N and the size of the alphabet.

			Value of N		
Dataset	Size of sequence dataset	Alphabet size	2	17	42
C2H2	26740	20 (protein)	0.084s	0.063s	0.061s
RG	27663	20 (protein)	0.112s	0.100s	0.100s
TOP	92226	20 (protein)	0.400s	0.331s	0.320s
Intron	81099	4 (DNA)	0.304s	0.260s	0.250s
TE	786738	4 (DNA)	5.958s	3.199s	2.296s

Tab 1: Time performance of N-local decoding algorithm.

1.3 Local decoding into segments

The information carried by the classes of the N-local decoding can be redundant. As a first step towards reducing this redundancy, we define *local segments of the N-local decoding*, which form a coarser partition of the sites in Σ that carries the same information.

Let R be the binary relation

$$R = \{(\sigma, \sigma+1) \in \Sigma \times \Sigma \,|\, Card([\sigma]_N) > 1 \text{ and } [\sigma+1]_N = [\sigma]_N + 1\},$$

where $[\sigma]_N + 1$ means the set obtained by taking the follower of every element in $[\sigma]_N$. The classes of the equivalence relation $\equiv_N$ spanned by R are called the *N-segments of GD-classes* of S. For any $\sigma \in \Sigma$, let $\Pi(\sigma)$ be the unique N-segment of GD-classes containing σ.

In other words, the N-segments of *GD*-classes are the common subwords of the sequences in S rewritten in the new extended alphabet of *GD*-classes.

1.4 Clustering of sequences

The N-local decoding thus described allows to construct groups of sequences in the following way.

Consider for a while a sequence s as the set of its positions $\sigma = (s, j)$ for $1 \leqslant j \leqslant |s|$. Given an integer N, and an equivalence class $\gamma \in P_N$, let

$$\mathcal{S}(\gamma) = \{s \in S \,|\, s \cap \gamma \neq \emptyset\}.$$

The set $S(\gamma)$ is the set of sequences that contain an element of the equivalence class γ.

The *groups of sequences clustered by* $\sim_N$ is the image $\mathcal{G}_N$ of the map $\mathcal{S} : P_N \longrightarrow \mathcal{P}(S), \gamma \longmapsto S(\gamma)$.

Conversely, for a subset $G \in \mathcal{G}_N$, let $\mathcal{S}^{-1}(G)$ be the *GD-class profile (of order N) of the subset G*. There is an obvious bijection between subsets clustered by $\sim_N$, and GD-class profiles of order N (which are all disjoint by construction).

1.5 From the suffix tree to the partitions tree

Let Π_Σ be the complete lattice of partitions of Σ endowed with the partial ordering $\leqslant$, and $\Delta = \{(\sigma,\sigma) \,|\, \sigma \in \Sigma\}$ be the minimum of Π_Σ.

Lemma 2.
For all $N \geqslant 0$, the partitions P_N satisfy $P_{N+1} \leqslant P_N$.

Proof. For all sites σ and σ', we have $\sigma \simeq_{N+1} \sigma' \Rightarrow \sigma \simeq_N \sigma'$. Therefore $[\sigma]_{N+1} \subset [\sigma]_N$. □

The partitions P_N thus naturally give rise to a tree $Partree(S)$ whose nodes of depth N are the classes of the relation $\sim_N$, and whose edges only connect a node γ of depth N to a node α of depth $N+1$ if $\alpha \subset \gamma$. As N reaches a certain value, the partition P_N reduces to Δ. Since a singleton of P_N does not carry any similarity information, we will assume that we have pruned $Partree(S)$ of all its singletons. In the same way, a valency 2 node of depth N, which corresponds to a class $[\sigma]_N$ that does not ramify at level $N+1$, that is such that $[\sigma]_{N+1} = [\sigma]_N$, will be suppressed from the tree, and the corresponding appearing edge, weighted by the number of suppressed nodes plus one. The set of *significant GD-classes* is the set

$$P_{sig} = \{\gamma = [\sigma]_N \in P_N \,|\, [\sigma]_{N+1} \neq [\sigma]_N, N \in \mathbb{N}\}.$$

For a given site $\sigma \in \Sigma$, the significant levels are the

$$N(\sigma) = \{k \in \mathbb{N} \,|\, [\sigma]_k \neq [\sigma]_{k+1}\},$$

and let $N_{max}(\sigma) = \max_{k \in N(\sigma)} k$. This last quantity is useful to construct local *motifs* (see section 2.1.3).

2 Results

We have applied some of these methods on two protein datasets, TOP and RG. The first is composed of all the 124 topoisomerases that have been sequenced up to march 2007, and the latter consists of 23 bacterial and archaeal reverse gyrases.

These 124 topoisomerase sequences are between 553 and 1067 residues long, and have 10 active sites (Forterre *et al.* [4]) as discovered by structural analysis. The second dataset studied by Nadal *et al.* [5] is composed of those 23 sequences among the previous ones that have an extension (the *reverse gyrase*) on the *N*terminal

side and who contain therefore 5 further catalytic sites. Both kinds of sequences are essential to DNA replication, and are therefore considered to be at the same time very ancient and conserved. However, they show sufficient variations in primary sequence to make both phylogenetic classification and multiple alignment (*e. g.* with ClustalW) of these sequences remain quite unreliable.

2.1 Clustering of sequences

All N put together, the total number of clusters of sequences amounts to 7870, which may seem a lot, but has to be compared with the 2^{124} possible subsets of the sequence dataset. An analysis of the indegree distribution of nodes of the Hasse diagram of this set of subsets (*i. e.* the directed graph on the set of subsets whose edges indicate minimal inclusion relations, (and layered by cardinality)) reveals some preferred associations between sequences, which roughly matches the known taxonomy, but has some interesting outliers.

We will here give a few examples of results obtained from N-local decoding of these sets of sequences.

2.1.1 Example 1: Correction of annotation

In the dataset TOP, *Thermoplasma Acidophilum* is annotated as a thermophile bacterium. The clustering resulting from the N-decoding consistently clusters this microbial sequence with Archaea. After curation of the databases, it turns out that the annotation was incorrect, and that Th. Ac. is indeed a thermophile Archaeum.

2.1.2 Example 2: Horizontal transfer hypothesis

The topoisomerase IA of *Arabidopsis thaliana* (AthT1) shows a clear preference towards bacterial sequences, and in particular with *Rickettsia prowazekii* (RprT1), an α-proteobacterium, with which it clusters up to $N = 20$.

```
RprT1_M_B_Rp  KEVIPNKHFTEPPPRYSEASLVKKLEELGIGRPSTYASILSVLQDRKYVALEKKRFIP
AthT1_M_E_At  GEVELKQHHTQHPPRYSEGSLVKKLEELGIGRPSTYASIFRVLQHRKYVTIKNRVLYP
```

Tab 2: Longest segment characterising the group *Arabidopsis thaliana*, *Rickettsia prowazekii*

This observation is consistent with an assumed mitochondrial origin of topoisomerase IA of *Arabidopsis thaliana* ([6]).

2.1.3 *Motif* detection

The most interesting feature of the N-local decoding is that, despite being an exact word matching combinatorial method, it manages to capture (mainly by virtue of the transitive closure mechanism) letter substitutions which are biologically meaningful. In the following example, we see how a single GD-class at $N = 7$, which is moreover the only invariant amino-acid for the whole pattern, reveals an active site with all its variations among the sequences. The following picture has been constructed by aligning the sites decoded by the same GD-class $P4$ for $N = 7$, and extending the window around a site σ to the width given by $N_{max}(\sigma)$ (see 1.5), in order to show all the relevant information carried around this site.

```
GD-class P4 (N=7)
Sequence name position
SsoR2_T_A_Ss    99    ---------------LASNQSFTMSA P TGLGKTTTLMT---------------
ApeR1_T_A_Ap   109    ----------------ARGDSFSIIA P TGVGKTTFGA-----------------
PkoRG_T_A_Pk    89    --------QRTWVKRLLKGRSFSIIA P TGMGKSTFGAFMAVWHAL--------
ApeR2_T_A_Ap   106    ------------------GDSFAIIA P TGVGKSTL-------------------
NeqRG_T_A_Ne    90    ---------------AYKGYSFSIIA P TGMGKTTFALV----------------
TpeRG_T_A_Tp   110    -------------------NTSFVILA P TGVGKTVF-------------------
TmaRG_T_B_Tm    90    ---------------IVQGKSFTMVA P TGVGKTTFGMM----------------
StoR1_T_A_St    93    ---------------LAKSESFSLSA P TGLGKTTTLLV----------------
AfuRG_T_A_Af    72    -------------------ESFAATA P TGVGKTS--------------------
PaeRG_T_A_Pa    98    ------------------GKSFAIVA P TGSGKTTF-------------------
AaeR2_T_B_Aa   106    --------------RVFMNQSFAIVA P TGVGKTTFGLVM---------------
PabRG_T_A_Pa    81    TGFRFWSAQRTWVKRILRGKSFSIIA P TGMGKSTFGAFISIYFAIKGKRSYIV
PfuRG_T_A_Pf    88    -------AQRSWVKRIIKGKSFSIIA P TGMGKSTFGAFMSIYFALK-------
AaeR1_T_B_Aa    86    ----------------VFLGRSFAMLA P TGVGKTTFGLS----------------
StoR2_T_A_St    91    ----------SWIIRVLRKESFAIIA P PGLGKTTFGIITSLYF-----------
SsoR1_T_A_Ss    92    -------PQRSWTIRFLRGESFAIIA P PGLGKTTFGLIMSLYNATR-------
MjaRG_T_A_Mj    95    ---------------------SFSIVV P TGVGKS---------------------
TteRG_T_B_Tt    90    --------------RLLLSKSFTLVA P TGVGKTTFGLIS---------------
SacR2_T_A_Sa    95    -----------------RGESFSLSA P TGVGKTTTL------------------
TthR1_T_B_Tt    81    ----------------VQGRSFAMLA P TGIGKTTFGL-----------------
PhoRG_T_A_Ph    81    TGFKFWSAQRTWVKRIIRGKSFSIIA P TGMGKSTFGAFISIYFATKGKKSYIV
SacR1_T_A_Sa    91    -------PQKSWIYRLLSGESFAIIA P PGLGKTTFGLISSIYLYLR-------
```

All of the 15 *motifs* of the reverse gyrases and 10 of the full topoisomerase set are found by our technique. The RG *motifs* can also be discovered by specially dedicated HMM-based tool such as MEME [1]. The comparison cannot however be extended to the TOP dataset, since it proves too large to be handled by MEME.

3 Conclusion

The N-local decoding, despite being a purely combinatorial procedure, exclusively based on primary sequence data, proves to capture biologically significant similarities between sequences, and provides a basis for an effective *ab initio*, unsupervised classification method for biological sequences, which can reasonably challenge established bioinformatic methodologies such as multiple alignment or HMM-based tools. Those characteristics of this method which have often been pinpointed as apparent drawbacks of the method, namely the choice of the parameter N, and the size of the resulting enriched alphabet can be in fact viewed, as we hope to have convinced the reader of this paper, as an element for a deeper application of these techniques, and will be the subject of more detailed work in the near future.

References

[1] T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28–36, AAAI Press, Menlo Park, California, 1994.

[2] G. Didier, M. Pupin, I. Laprevotte and A. Hnaut. Local decoding of sequences and alignment-free comparison. *J Comput Biol*. Oct;13(8):1465-76, 2006.

[3] G. Didier, L. Debomy, M. Pupin, M. Zhang, A. Grossmann, C. Devauchelle and I. Laprevotte. Comparing sequences without using alignments: application to HIV/SIV subtyping. *BMC Bioinformatics*. 2,8:1, 2007.

[4] Forterre, P., S. Gribaldo, D. Gadelle, and M. C. Serre. Origin and evolution of DNA topoisomerases. *Biochimie* (under press), 2007.

[5] Jaxel, C., C. Bouthier de la Tour, M. Duguet, and M. Nadal. Reverse gyrase gene from Sulfolobus shibatae B12: gene structure, transcription unit and comparative sequence analysis of the two domains. *Nucleic Acids Res* 24:4668-75, 1996.

[6] Nadal, M. Reverse gyrase: An insight into the role of DNA-topoisomerases. *Biochimie* (under press), 2007.

[7] Ukkonen, E. On-line construction of suffix-trees. *Algorithmica* 14 (1995), 249-260