Università di Pisa
Dipartimento di Informatica

Technical Report: TR-05-

Implicit and Explicit Representation of Approximated Motifs

Nadia Pisanti
Henry Soldano
Mathilde Capentier
Joel Pothier

September 17, 2005

ADDRESS: via F. Buonarroti 2, 56127 Pisa, Italy. TEL: +39 050 2212700 FAX: +39 050 2212726

Implicit and Explicit Representation of Approximated Motifs

Nadia Pisanti [*] Henry Soldano [†] Mathilde Capentier [‡]
Joel Pothier [§]

September 17, 2005

Abstract

Detecting repeated 3D protein substructures has become a new crucial frontier in motifs inference. In [7] we have suggested a possible solution to this problem by means of a new framework in which the repeated pattern is required to be conserved also in terms of relations between its position pairs. In our application these relations are the distances between α-carbons of amino acids in 3D proteins structures, thus leading to a *structural consensus* as well. In this paper we motivate some complexity issues claimed (and assumed, but not proved) in [7] concerning inclusion tests between occurrences of repeated motifs. These inclusion tests are performed during the motifs inference in *KMRoverlapR* (presented in [7]), but also within other motifs inference tools such as *KMRC* ([9]). These involve alternative representations of motifs, for which we also prove here some interesting properties concerning pattern matching issues. We conclude this contribution with a few tests on cytochrome P450 protein structures.

1 Introduction

Finding repeated subsequences and substructures in biological (resp. sequential and structural) data is having growing importance for various different applications in molecular biology. Among them we can mention the detection of trasncription factors binding sites as repeated gapped motifs in the upstream regions preceeding genes, or the prediction of RNA secondary structures as complementary reversed repeated subsequences, the detection of common fragments

---

[*] Dipartimento di Informatica, University of Pisa, Italy and Laboratoire d'Informatique de Paris Nord, University of Paris 13, France.

[†] Laboratoire d'Informatique de Paris Nord, University of Paris 13, France and Atelier de Bioinformatique, University of Paris 6, France.

[‡] Atelier de Bioinformatique, University of Paris 6, France.

[§] Atelier de Bioinformatique, University of Paris 6, France.

of genomic sequences as a starting point of measuring genomic distances, etc. In this paper we focus on yet another biological application, that is the detection of common substructures in 3D proteins. In [7] we have designed an algorithm for the inference of repeated motifs under the new framework of *relational* motifs which results particularly suitable for this purpose.

Motifs inference in biological applications requires a certain degree of approximation in establishing whether a biological object is basically the same as another one. For this reason, the possibly huge size of solutions in the search space makes the algorithmical solution tricky. It is very difficult to find the right balance between the sensitivity of a motif inference tool and its efficiency when an exhaustive algoritmical approach is suited. Most of the difficulty comes from the unavoidable noise of biological data which causes an explosion of intermediate candidates (typically, shorter motifs to be extended or composed to make longer ones). Hence, it is very important that the inference tool offers a way to refine the query in order to minimize this noise. For this purpose, we have designed *KMRoverlapR* that detects repeated motifs that are approximated because they are patterns defined on a input degenerate alphabet, and they are also required to be conserved in terms of relations between each pair of positions of the consensus sequence. In our application to 3D proteins, the input sequences are amino acid sequences enriched with the information, per each pair of positions that are at most at a distance of k letters, of the distance between the corresponding α-carbons in the 3D structure. Moreover, the amino acids are grouped into possibly overlapping subgroups that somehow represent similar physical and chemical characteristics. Finally, also for the distance between the α-carbons, it is given a set of possibly overlapping ranges of values. A relational k-pattern is a k-long sequence of the groups of amino acids among those given above, with the ranges of its $k(k-1)/2$ distances between the α-carbons, in the 3D structure, of each pair of distinct positions. A pattern *occurs* in the input sequence if the latter contains a k-long fragment where in all positions the amino acid belongs to the corresponding group of the pattern, and each pair of them is at a distance in the 3D structures that fits the ranges of the distances required by the pattern's relations. Given a quorum q, the goal is to detect all (relational) k-patterns that occur at least q times, and that we will name (relational) k-motifs.

In [7] we have introduced a linear time algorithm for the inference of repeated relational k-motifs. Thanks to some properties proved in [7], the algorithm guarantees a complete and correct inference avoiding to have to list all candidates in intermediate steps. This is achieved by means of an implicit representation that only uses the *extent* of a motif (*i.e.* the complete set of its occurrences), and thanks to some sufficient conditions that allow to keep only distinct extents all along the computation. In this paper we address some issues concerning this implicit representation and its possible alternatives for some specific purposes. In particular, we will discuss properties of a couple of explicit representation for possible post-processing of the inferred motifs, and we will motivate some complexity issues claimed (but not proved) in [7] that involve the representation of the motifs.

2 Previous Work: KMRC and KMRoverlapR

In [7] we have introduced a tool for inferring approximatively repeated relational motifs. The framework of relational motifs is very powerful in that it may allow refined queries thus leading to sensible and, at the same time, efficient inference of repeated substructures. In [7] we have given motivations for choosing a *KMR*-like approach ([6]) when relations are taken into account [1] which, in turn, leads to the choice of a degenerate alphabet to express the approximation. The degenerate alphabet to be used to describe the motif is explicitly given as an input parameter (a second degenerate alphabet is possibly also given for the relations in *KMRoverlapR*) under the form of a cover G over the alphabet Σ of the input sequence. Each element of this cover is a subset of Σ and we will denote these elements as *groups*. We denote with *degeneracy* g the maximum number of distinct groups to which a letter belongs to. A motif is thus seen as a sequence of groups $C_1 \ldots C_k$ that occurs in the input sequence at position p whenever there is a sequence of letters $\sigma_1 \ldots \sigma_k$ such that $\sigma_j \in C_j \in G$ for $1 \leq j \leq k$. The inference algorithm of *KMRoverlapR* ([7]) is a suitable extension of the *KMRC* one ([9]). In [7] we have addressed some issues that raised specifically for relational motifs in *KMRoverlapR*. Nevertheless, many properties concerning the compact representation of the motifs by means of their extents, the filtering of maximal motifs, as well as complexity issues, are actually shared by the two tools. Among these, there are the issues discussed in this paper. For this reason, and since they can be straightforwardly extended to the case of relations, in what follows we will refer to motifs without relations in order to simplify the notation. We will denote with *k-motif* a motif of length k.

A common feature of *KMRC* and *KMRoverlapR* is the restriction to *maximal* motifs. A maximal k-motif is a motif whose complete list of occurrences, that we will name *extent*, is not properly included into the extent of another k-motif. It is a *duplication* if it is equal. In [9] an upper bound of the total size of the extents of k-motifs has been proved, and in [7] its natural extension to relational motifs is shown. This bound is theoretically the same whether or not we restrict to maximal and non duplicated motifs only. Nevertheless, in practice we observed noticeable ratios between their number (see [7] for details). This, and the fact that maximal motifs of a fixed length suffices to infer all distinct maximal motifs of greater length, motivates the elimination, at each intermediate step, of all non maximal (or duplicated) extents. This requires an exhaustive inclusion test between all pairs of candidate motifs, which becomes actually the bottleneck of the computation.

Omitting variants and specific features due to the introduction of relations,

---

[1] That is, an *in width* inference of motifs in the sense that all motifs of length ℓ are inferred before any motif of length $> \ell$. The complementary choice is an *in depth* approach like [8] where each single candidate is extended as long as it satisfies the requirements. In general, when the motif is represented as a consensus pattern, the *in depth* results a better choice ([4]), but in [7] we have shown that with relations the things change.

the inference algorithm we refer to can be summarized in the following steps where we assume that we seek maximal k-motifs, that are k-long words of the alphabet of the groups, that occur at least q times in an input sequence s of length n.

1. Compute extents of each group, *i.e.* compute extents of ($\ell = 1$)-motifs.
2. **while** $\ell < k$ **do**
   (a) Compute extents of ($\ell + d$)-motifs from those of ℓ-motifs; $\ell := \ell + d$;
   (b) Eliminate extents containing less than $< q$ occurrences.
   (c) Eliminate extents that are included into others.
3. Output all left extents (that is, all maximal k-motifs).

In other words, it is possible to perform the inference keeping only the extents of the motifs, that are ordered subsets of $\{1, \ldots, n\}$. On these the most involving operation we do is the $2(c)$ above consisting in the detection of extents that are equal to or included into others. We will denote step $2(c)$ as *Inclusion Test*.

3 Representation of Maximal Motifs

3.1 Implicit Representation with Occurrences Lists

The set of distinct patterns of length k can obviously be as big as the set of different k-long words on the alphabet G of the groups, which has size $|G|^k$. For example, in the simple (although improbable) sequence $\overline{\sigma}^n$, if $\overline{\sigma} \in \Sigma$ occurs in all groups of G, then we have that every string in G^k is a k-motif for $1 \leq k \leq n-1$ (for any quorum $1 \leq q \leq n-k+1$). Hence, the upper bound mentioned above happens to be tight and therefore an explicit representation of all motifs of a given length is unfeasible. On the other hand, the exponential number of motifs shown above can be represented by an unique extent $X = \{1, 2, \ldots, n-k+1\}$ and a length k, (that is in linear space). This is, intuitively, the motivation of why the algorithm of *KMRC* and *KMRoverlapR* actually deal with extents only. In fact, the above mentioned motifs of the sequence $\overline{\sigma}^n$ can all be represented by an unique extent because they are all maximal duplications of each other. We observed (see [7]) a ratio in $O(10^3)$ between the number of maximal motifs and the nonmaximal ones, when the latter are $O(10^5)$. Therefore, in practice the gain is noticeable.

3.2 Explicit Representation of Motifs

The implicit representation described in Section 3.1 allows a sensible speed up in the inference phase, and in particular it avoids an explosion of generated candidates. Nevertheless, for the purpose of describing the output, a more explicit representation would be more suitable, in order to *visualize* the actual motifs once their inference is performed. Moreover, as we will see in Section 4, also some complexity issues motivate a switch to an explicit representation already during the inference phase.

An extent X of a motif actually represents the following set of motifs.

$$M(X) = \{C_1 C_2 \cdots C_k \mid s[p+j-1] \in C_j,\ \forall\, 1 \leq j \leq k,\ \forall\, p \in X\}.$$

Moreover, we will denote with $M_j(X)$ the set of groups that are at position j in $M(X)$, that is $M_j(X) = \{C_j \mid C_1 \cdots C_j \cdots C_k \in M(X)\}$ for any $1 \leq j \leq k$.

Example In the simple sequence $s = abbbc$ with the cover $G = \{C_1 = \{a,b,z\}, C_2 = \{b,c,z\}, C_3 = \{x\}\}$, we have that the extent $X = \{1,2,3\}$ of a 3-motif represents the set of motifs $M(X) = \{C_1C_1C_2, C_1C_2C_2\}$ and that $M_2(X) = \{C_1, C_2\}$.

The motifs set $M(X)$ is itself an explicit representation, but it can results too redundant. There are more compact ways — still more explicit than the extents — to represent $M(X)$. We report here below two possibilities.

1. ***G*-representation.** A k-long sequence of intersections of groups of the cover G. For each position $1 \leq j \leq k$ we have

$$G_X[j] = \cap_{C_j \in M_j(X)}\, C_j.$$

2. **Σ-representation.** A k-long sequence of subsets of Σ, listing for each position $1 \leq j \leq k$ the set $\Sigma_X[j]$ of letters occurring at position j in the occurrences. Formally:

$$\Sigma_X[j] = \cup_{p \in X}\, s[p+j-1].$$

Example Let us consider again the input text $s = abbbc$, the cover $C_1 = \{a,b,z\}, C_2 = \{b,c,z\}, C_3 = \{x\}$, and the extent $X = \{1,2,3\}$ representing $M(X) = \{C_1C_1C_2, C_1C_2C_2\}$. Its G-representation is $G_X = C_1(C_1 \cap C_2)C_2 = \{a,b,z\}(\{a,b,z\} \cap \{b,c,z\})\{b,c,z\} = \{a,b,z\}\{b,z\}\{b,c,z\}$, and the Σ-representation is $\Sigma_X = \{a,b\}\{b\}\{b,c\}$.

There are other possible representations, among which we can mention position specific scoring matrices (that is, a $|\Sigma| \times k$ table where for each $\sigma \in \Sigma$ and for each $1 \leq j \leq k$ we report the number of times the letter σ occurs at position j in an occurrence of the motifs set) as well as a variant containing the same information in terms of groups. The information of the distribution of the letters/groups could also be added in the G- and Σ- representations by using multisets rather than simple sets. These could be suitable for applications in which the statistics of the distributions of the letters is useful, and it can even result efficient for small size alphabets. These two conditions may hold for consensus sequences in DNA or RNA sequences. Nevertheless, in this paper we will concentrate on the two G- and Σ- representation formalized above because, as we will see in Section 4, they raise interesting complexity results for some crucial steps in the inference of motifs performed by *KMRoverlapR*.

The G- and Σ- representations are somehow related in that they eventually both display the motif as a sequence of subsets of Σ. Notice that, even if the Σ-representation is somehow independent from G, this latter has driven the inference and hence the resulting output. As a consequence, there are a few relations among the G-representation, the Σ-representation, and the cover G.

Lemma 1 *For all* $1 \leq j \leq k$, $\Sigma_X[j]$ *is a subset of at least one group of* G.

Proof: By definition, X is the extent of at least one motif $C_1 \cdots C_J \cdots C_k$, and hence $s[p+j-1] \in C_j \ \forall p \in X$ and thus $\Sigma_X[j] \subseteq C_j$. $\bowtie$

Actually, for the very same reason we have that $\Sigma_X[j] \subseteq C_j$ for each distinct C_j whose intersection is $G_X[j]$, which leads to the following result that is a direct consequence of Lemma 1.

Proposition 1 *For all* $1 \leq j \leq k$, $\Sigma_X[j] \subseteq G_X[j]$.

Depending on the application, the output may require that also the cover G is given in order to reconstruct lost information. We will discuss in Section 3.3 some pattern matching issues resulting from some loss of information in the different representations.

3.3 Searching occurrences in a new text

One of the possible need of an explicit representation of inferred motifs is the post-processing of such data. For example, in biological applications, the patterns resulting from motifs inference are often object of successive queries in pattern matching in order to check their occurrences in a new text. It is clear that if we want to search occurrences of an inferred motif into a new text, we need to process the extent and write an explicit representation of *what* we want to search. In this section we address some issues concerning such queries considering the distinct possible representations we suggested in this paper.

Example Let us consider the same cover $G = \{C_1 = \{a,b,z\}, C_2 = \{b,c,z\}, C_3 = \{x\}\}$, $k = 3$, and the input string s as in the previous example. For the extent $\{1,2,3\}$, the G-representation is $G_X = \{a,b,z\}\{b,z\}\{b,c,z\}$, and the Σ-representation is $\Sigma_X = \{a,b\}\{b\}\{b,c\}$. Let us now consider the new text $s' = \underline{azc}xxx\underline{aab}xxx\underline{abc}$ and, in particular, the underlined substrings of length 3 occurring, respectively, at positions $1, 7$, and 13. If we search the G-represented pattern $\{a,b,z\}\{b,z\}\{b,c,z\}$, we would only find the occurrences 1 and 13. Moreover, searching the Σ-represented $\{a,b\}\{b\}\{b,c\}$ we get position 13 only.

Nevertheless, notice that an *ex-novo* inference of maximal 3-motifs occurring at least three times in s' and written in the alphabet of the cover G would result in the extent $\{1, 7, 13\}$ representing the k-motif $\{C_1\}\{C_1\}\{C_2\}$.

The example has shown that the two representations behave in general differently in possible post-inference text search of a motif. Moreover, they both miss occurrences with respect to a possible ex-novo inference with the same parameters. However, depending from the application, it can be that what one actually wants to find is not the complete set of occurrences as if the motif were inferred from scratch, but rather the possible position where a specific instance of it occurs. For example, assume that we have inferred an over represented fragment in a set of 3D protein structures. Assume that for the spatial distance we have been using an alphabet that groups a range of possible distances in the interval $[d_{min}, d_{max}]$, but that we have detected a frequent substructure having always basically the same distance $\overline{d} \in [d_{min}, d_{max}]$ between two specific positions. It is reasonable to think that after such inference one wants to search this specific observed pattern. In this sense explicit representations with loss of information such as the two above can still result as valid.

3.4 Complexity Issues of Explicit Representation

In this section we analyse time and space complexity of computing and storing the two different explicit representations.

Computing the G-representation requires an exhaustive search in all positions of all occurrences of the motifs X represents. And per each one of them the degeneracy of G has to be taken into account as well. We assume that we have a vector V containing, for each $1 \leq i \leq n$, the set $V[i]$ of groups occurring at position i of the input sequence[2]. By definition of V and $G_X[j]$, we have that, for all $p \in X$, if $s[p+j-1] \in C_j$, then $C_j \in V[p+j-1]$. Nevertheless, for different $p \in X$ there are obviously different sets $V[p+j-1]$, each one being in general a superset of $M_j(X)$ and thus of $G_X[j]$. We have the following useful result.

Proposition 2 $G_X[j] = \cap_{p \in X} V[p+j-1]$.

Proof: We have that $G_X[j] \subseteq \cap_{p \in X} V[p+j-1]$ as a direct consequence of the fact that, $\forall\, p \in X$, if $s[p+j-1] \in C_j$ then $C_j \in V[p+j-1]$. For the opposite ($\cap_{p \in X} V[p+j-1] \subseteq G_X[j]$) let us fix j and consider any $C_j \in G$ such that $C_j \in \cap_{p \in X} V[p+j-1]$. By definition of V this means that $s[p+j-1] \in C_j$ $\forall p \in X$ and thus that $C_j \in M_j(X)$. Hence, $G_X[j]$ contains the intersection of all such C_j's proving the thesis. $\bowtie$

As a consequence of Proposition 2, $G_X[j]$ can be computed as the intersection of $|X|$ lists of groups, each one ($V[p+j-1]$) containing at most g

---

[2] This data structure is actually created both in *KMRC* and in *KMRoverlapR* and kept during the inference phase.

elements. Such lists are ordered and $\cap$ is associative, and thus it suffices to perform a linear visit to the lists to compute the intersection. Hence, given that $\sum_X |X| \le ng^k$,computing the G-representations of all the extents X of maximal k-motifs takes $\sum_X |X| \cdot k \in O(ng^k k)$ time in the worst case. The space complexity is also in $O(ng^k k)$.

The Σ-representation of all maximal k-motifs can be computed, for all extents X, and for all positions $1 \le j \le k$, by doing the union of $s[p+j-1] \; \forall \; p \in X$, which can result in at most $|\Sigma|$ elements, then taking $\sum_X |X| \cdot |\Sigma| \in O(ng^k|\Sigma|)$ time and space.

4 Inclusion Test with Explicit Representation

Both in *KMRC* and in *KMRoverlapR* the bottleneck is the motifs inference in the elimination of non-maximal motifs. This requires an exhaustive search in extents included into others, and the inefficiency is caused by the fact that all extents must be pairwise tested for a possible inclusion, each one of them can contain as many as n elements. This drawback could be avoided with an explicit representation of motifs because the comparison would be performed between objects of size at most n. In this section we show necessary and sufficient conditions on the explicit representations that correspond to inclusion among extents. Let us start with observing the following fact which is a direct consequence of the fact that extents inferred by *KMRC* and *KMRoverlapR* are the complete set of occurrences of one or more maximal motifs.

Fact 1 *Any maximal extent X extracted by* KMRC *or* KMRoverlapR *from a sequence s has the property that there exists no $\overline{p}$ such that $\overline{p} \notin X$ and $s[\overline{p}+j-1] \in \cup_{p\in X} s[p+j-1] \; \forall \; 1 \le j \le k$.*

We give now a necessary and sufficient condition to detect nonmaximal or duplicated motifs within the explicit representation. In what follows, we will say that $\Sigma_{X'} \subseteq \Sigma_X$ if $\Sigma_{X'}[j] \subseteq \Sigma_X[j] \; \forall \; 1 \le j \le k$, and similarly that $G_X \subseteq G_{X'}$ if $G_X[j] \subseteq G_{X'}[j] \; \forall \; 1 \le j \le k$. Let us start observing that $X' \subseteq X \iff M(X) \subseteq M(X')$ because adding positions p where a motif is required to occur can only decrease the set of motifs satisfying the condition.

Lemma 2 *Let X, X' be extents of k-motifs. We have that*

$$X' \subseteq X \iff \Sigma_{X'} \subseteq \Sigma_X.$$

Proof: If $X' \subseteq X$, then we have that $\Sigma_{X'}[j] = \cup_{p\in X'} s[p+j-1] \subseteq \cup_{p\in X} s[p+j-1] = \Sigma_X[j] \; \forall \; 1 \le j \le k$.
If $\Sigma_{X'} \subseteq \Sigma_X$ then we have that $\cup_{p\in X'} s[p+j-1] \subseteq \cup_{p\in X} s[p+j-1]$ for all $1 \le j \le k$. Hence it must be that $M(X) \subseteq M(X')$ and thus that $X' \subseteq X$. $\bowtie$

Actually, a slightly stronger result (althought not useful for the purpose of this section) than Lemma 2 holds. Namely, we have that $\Sigma_{X'} \subsetneq \Sigma_X \iff X' \subsetneq X$ because if $\exists\ j'$ and $\tilde{\sigma} \in \Sigma$ such that $\tilde{\sigma} \in (\Sigma_X[j'] \setminus \Sigma_{X'}[j'])$, then we have a position $\tilde{p} \in X$ such that $s[\tilde{p} + j - 1] = \tilde{\sigma} \notin \cup_{p \in X'} s[p + j' - 1]$, which implies that $\tilde{p} \notin X'$ and thus that $\tilde{p} \in (X \setminus X')$.

In terms of G-representation, we have a similar result.

Lemma 3 *Let X, X' be extents of k-motifs. We have that*

$$X' \subseteq X \iff G_{X'} \subseteq G_X.$$

Proof: If $X' \subseteq X$, then we have that $M(X) \subseteq M(X')$ and thus that $\forall\ 1 \leq j \leq k$ $G_{X'}[j] \subseteq G_X[j]$ because, in general, in $M_j(X')$ there are at least as many groups to intersect as in $M_j(X)$, and further intersections can only decrease the final set. These implications can be easily reversed in case of equality all over.
We still need to prove that $G_{X'} \subsetneq G_X \Rightarrow X' \subseteq X$. The hypothesis implies that there exists one or more j' such that $G_{X'}[j'] \subsetneq G_X[j']$ (and in other positions $G_{X'} = G_X$). This means that $\cap_{C \in M_{j'}(X')} C \subsetneq \cap_{C \in M_{j'}(X)} C$, and hence that $M_{j'}(X) \subsetneq M_{j'}(X')$ and in general $M(X) \subseteq M(X')$, which implies $X' \subseteq X$.

$\bowtie$

Example Let us consider the string *xbxcxaxbxc* and the cover $C_1 = \{a, b\}, C_2 = \{b, c\}, C_3 = \{x\}$. We have that $X = \{2, 6, 8\}$ and $X' = \{2, 8\}$ are such that $X' \subseteq X$ and in fact $M(X) = \{C_1C_3C_2\}$ and $M(X') = \{C_1C_3C_2, C_2C_3C_2\}$. We have that $\Sigma_{X'} = \{b\}\{x\}\{c\} \subset \Sigma_X = \{a, b\}\{x\}\{b, c\}$ and $G_{X'} = \{C_1 \cap C_2\}\{C_3\}\{C_2\} \subset G_{X'} = \{C_1\}\{C_3\}\{C_2\}$.

Lemma 3 and 2 allow to conceive a different way to perform inclusion tests in order to detect and discard duplicated and nonmaximal k-motifs. Besides the use of explicit representation, the idea of this fast inclusion test is to compare only extents that share a position. This is done by ranging over all positions of the input strings and for each position i we only compare pairs of k-motifs that both occur at position i.

Proposition 3 *Detecting nonmaximal and duplicated extents can be done in $O(ng^{2k}k\ell)$ time, where $\ell = min\{g, |\Sigma|\}$.*

Proof: In order to detect pairs of extents that are equal or included one into the other, the necessary and sufficient conditions of Lemma 3 and 2 allow to compare explicit representations.
Using the G-representation requires to check, for each position in the input sequence (there are n of them), per each pair of motifs both occurring in that position (there are g^{2k} of them), and per each one of their k positions, whether the two ordered lists of at most g elements are included one into the other. The resulting time complexity is in $O(ng^{2k}kg) = O(ng^{2k+1}k)$.
Similarly, with the G-representation we should do, per each pair of motifs and

per each one of their k positions, an inclusion test between two ordered sets of size at most $|\Sigma|$. As a result, we need $O(ng^{2k}k|\Sigma|)$ time. $\bowtie$

Of course, if at each intermediate step of the inference the inclusion step is performed on an explicit representation of motifs, then this latter has to be computed and the cost of this computation must then be taken into account. Nevertheless, this still results into a worst case complexity in $O(ng^{2k}k|\Sigma|^2)$ which is an improvement over the time cost in $O(n^2g^{2k})$ of the inclusion tests performed over the extents because $k, g, |\Sigma| << n$. Indeed, the cost of inclusion test, and hence of the whole inference, becomes linear in the size of the input sequence, thus eliminating the drawback of the quadratic time complexity with respect to the size n of the input sequence.

5 Applications to 3D protein structures

As mentioned in the section 1, relational motifs can represent structural motifs in 3D proteins structures. For this purpose the relation between two points x_p and x_q is obtained by discretizing the euclidian distance $d(x_q, x_p)$. Relational motifs represent then geometrical motifs in a multidimensional space, i.e. motifs whose occurrences are insensitive to translations and rotations. Such internal distances between atoms were first used to search structural motifs in the general context where all the atoms of the protein are considered here in [1] and, using a tolerance as here, in [2]. When only considering the α-Carbons in the 3D structure of the protein we obtain a sequence of points in a 3D space.

Here we consider that a prior discretization of the distances has been performed and that relations are denoted as positive integers. We consider a set of relational groups $\{R_j = \{j, ..., j+\delta\}\}$ where δ represents a tolerance level: two discretized distances $d(x_p, x_q)$ and $d(x_{p'}, x_{q'})$ belong to the same group whenever $|d(x_p, x_q) - d(x_{p'}, x_{q'})| \leq \delta$. Note that as a consequence we have that the degeneracy of the relations's alphabet is $\delta + 1$.

Hereunder we give an example of the results obtained when searching a structural pattern repeated in the backbone of several proteins. We chose to study structures of the cytochrome P450 multigenic superfamily (CYP, P450). They are heme-thiolate proteins involved in many oxidations of hydrophobic substrates. The substrates are steroid hormones, extracellular fatty acids signaling molecules and vitamins but also exogenous substrates as drugs or environmental pollutants (see [3] for an historical review). These P450s can be found in many living beings: bacteria, yeast, fungi, plants, insects, fishes and mammals. They have been widely investigated notably because of their role in drugs degradation. Their amino-acids primary sequences are dissimilar in spite of their structural similarities.

We chose five cytochrome P450 structures: four from bacteria (PDB codes 4CP4, 1CPT, 2HPD chain B and 3CPP) and one from fungi (PDB code 1ROM). Note that here the algorithm searches for patterns that occur at least q times in a set of m protein structures. Here $m = q = 5$. The distances between α-Carbon

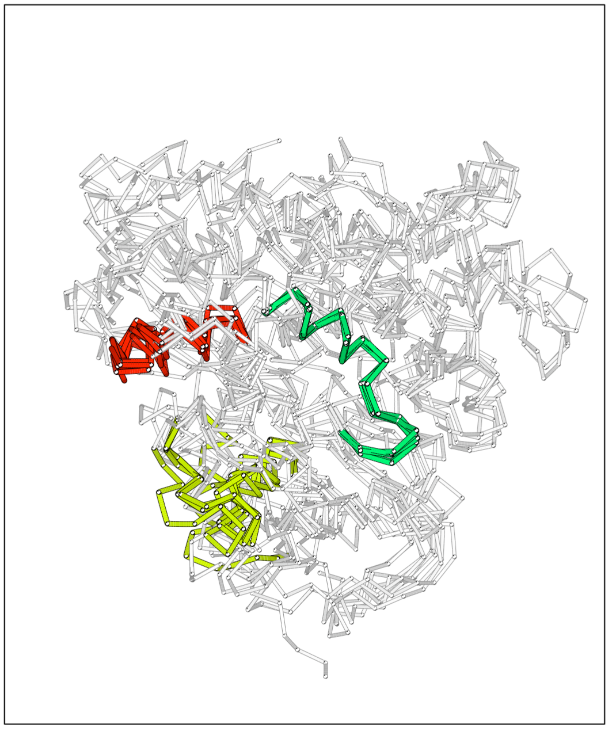
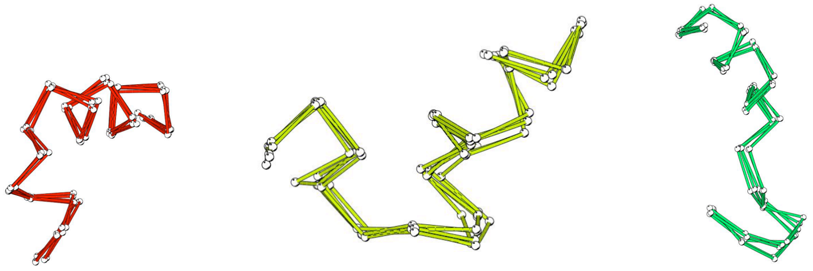


Figure 1: Structural relational motifs of length 18 for five cytochromes P450; their PDB codes are: 4CP4 ,1ROM, 1CPT, 2HPD chain B and 3CPP. This is the maximum length reached using 0.5Å-long intervals, a tolerance level $\delta = 2$ (shorter motifs are not shown). Protein backbones are in grey and motifs are colored and thicker; only α carbons are represented (white small balls) and we trace pseudo-bonds between them (scale 3.8Å between two consecutive C_α). As these locally matching substructures could have slid in one structure with respect to another structure, they may not be aligned all at once. Here structures are aligned according to the green motif.

are discretized using 0.5Å-long intervals, with a tolerance level $\delta = 2$ (so that the degeneracy is 3). The average length of the sequences considered is about 400 residues. There are 10 relational groups (amongst 40) representing distances



(a) (b) (c)

Figure 2: Occurences of the three structural relational motifs of 18 residues. All of them are also found in [5]. They are composed of a coil and a part of α helix. The first motif is the beginning of a very long one going through P450. The last one includes a well conserved Cysteine which bind the heme (7th residue of the motif).

actually appearing in the sequences. We represent hereunder in Figure 1 the occurrences of the longest structural motifs (k=18) found on the 5 proteins. Such motifs were previously identified on 3 of these proteins [5]. We also show in the table 1 a partial Σ-representation (because only the relations are given) of a motif. As the motif is relational, the Σ-representation is a set of $k(k-1)/2$ constraints, each one expressed as a distance interval δ_{ij}: in order to find an occurrence of this motif at position p in the backbone of a protein, for any pair of positions (i,j) in the motif, the distance between the $p+i^{th}$ and $p+j^{th}$ $\alpha - Carbons$ of the protein has to belong to δ_{ij}.

6 Conclusion

The method discussed here has been applied to the problem of matching substructures in several protein structures, and the results have been satisfying. As a test case, the matching substructures problem allowed direct (visual) inspection of the fitness of the algorithm, as similar relational motifs are 3D-matching substructures. In this case, the groups of relations - to be considered in the building of relational motifs - are computed as ranges of distances. More generally, relational motifs would be of interest in the biological sequences field, as not only the letters (residues) of the sequences are important, but also relations between some pairs of residues composing a relevant biological motif. And these relations can be not computed but assessed. For a simple example, one can cite amphiphatic helices comparison: in such an helix, a residue has more or less the same hydrophobic index than its neighbours have, and has an hydrophobic

Table 1: Internal distance intervals (Å) computed for the six first residues of the third motif (green one on figures 1 and 2(c)). The distance between two consecutive α-carbons is always nearby 3.8Å, therefore matrix diagonal values are from 3.7Å to 3.9Å. On figure 2(c) the first residue is at the bottom and, due to the coil, α-Carbons 1 and 2 are closer to the α-Carbon 6 than to the α-Carbons 3, 4 and 5. As distances are discretized using 0.5Å-long intervals and the tolerance level is $\delta = 2$, distance differences are always less than 1.5Å.

Residue index	2	3	4	5	6	7
1	[3.8 − 3.8]	[6.9 − 7.2]	[8.7 − 9.1]	[6.7 − 7.4]	[4.4 − 5.0]	[6.2 − 6.5]
2		[3.7 − 3.8]	[6.6 − 7.0]	[5.9 − 7.0]	[4.4 − 5.2]	[7.8 − 8.2]
3			[3.8 − 3.8]	[5.4 − 5.7]	[5.3 − 6.0]	[9.1 − 9.7]
4				[3.8 − 3.9]	[5.5 − 5.8]	[8.6 − 9.1]
5					[3.8 − 3.9]	[6.1 − 6.7]
6						[3.8 − 3.8]

index opposite to the residues located at position +4 or -4. The simple relations used here would be ”to have similar hydrophobic index” or ”to have opposite hydrophobic index”. Indeed, relations to be used in the biological sequences field can be much complex than those used in this example.

References

[1] C. W. Crandell and D. H. Smith. Computer-assisted examination of compounds for common three-dimensional substructures. *Journal of Chemical Information and Computer Sciences*, 23(4):186–197, 1983.

[2] V. Escalier, J. Pothier, H. Soldano, and A. Viari. Pairwide and multiple identification of three-dimensional common substructures in proteins. *Journal of Computational Biology*, 5(1):41–56, 1998.

[3] R. W. Estabrook. A passion for p450s (rememberances of the early history of research on cytochrome p450). *Drug Metab Dispos*, 31(12):1461–73, 2003. 0090-9556 Historical Article Journal Article Review Review, Tutorial.

[4] M. Tompa et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144, 2005.

[5] P. Jean, J. Pothier, P. M. Dansette, D. Mansuy, and A. Viari. Automated multiple analysis of protein structures: application to homology modeling of cytochromes p450. *Proteins*, 28(3):388–404., 1997.

[6] R.M. Karp, R.E. Miller, and A.L. Rosenberg. Rapid identification of repated patterns in strings, trees and arrays. In *Fourth ACM Symposium on Theory of Computing*, pages 125–136, 1972.

[7] N. Pisanti, H. Soldano, and M. Carpentier. Incremental Inference of Relational Motifs with a Degenerate Alphabet. In *Combinatorial Pattern Matching (CPM)*, pages 229–240. Springer-Verlag, 2005. LNCS 3537.

[8] M.-F. Sagot. Spelling approximate repeated or common motifs using a suffix tree. In *Latin American Theoretical INformatics symposium (LATIN)*, pages 111–127. Springer-Verlag, 1998. LNCS 1380.

[9] H. Soldano, A. Viari, and M. Champesme. Searching for flexible repeated patterns using a non-transitive similarity relation. *Pattern Recognition Letters*, 16:243–246, 1995.