**DIMACS Technical Report 2001-02**
**January 2001**

# Multiple Sequence Alignment Using the Quasi-concave Function Optimization Based on the DIALIGN Combinatorial Structures

by

**Leonid Shvartser**

Ness A.T.Ltd
PO Box 58180
Tel Aviv 61581, Israel
Leonid.Shvartser@ness.com

**Casimir Kulikowski, Ilya Muchnik**

Department of Computer Science
Rutgers University
P.O. Box 8018, Piscataway, NJ 08855
kulikows@cs.rutgers.edu

---

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, Bellcore and NEC Research Institute.

DIMACS is an NSF Science and Technology Center, funded under contract STC-91-19999; and also receives support from the New Jersey Commission on Science and Technology.

# Abstract

Multiple sequence alignment is usually considered as an optimization problem, which has a statistical and a structural component. It is known that in the problem of protein sequence alignment a processed sample is too small and not representative in the statistical sense though this information can be sufficient if an appropriate structural model is used. In order to utilize this information a new structural description of the pairwise alignment results union has been developed. It is shown that if the structure is restored then Multiple Sequence Alignment is achieved. Introduced structure represents the set of local maximums of quasi-concave set function on a lower semi lattice, which in turn is a union of the set-theoretical intervals. This union is a set of the consistent subsets of diagonals, introduced by B. Morgenstern, A. Dress, and T. Werner (1996). Algorithm for local maximums search on proposed structure has been developed. It consists of an alternation of the Forward and Backward passes. The Backward pass in this algorithm is a rigorous while the Forward pass is based on heuristics. Multiple alignment of 5 protein sequences are used as an illustration of the proposed algorithm.

## 1. Introduction

Procedures of amino acid sequence alignment are basic methods for similarity analysis among proteins (generally speaking the same is true for DNA sequences, but in the paper we are interested in protein classification; this is the reason why we are focused on amino acid sequences). The analysis is used to predict structural and functional properties of proteins based on their primary structures, for their evolution relations, etc. [1-12, 14, 16-21]. There are two types of the alignment procedures: for pair-wise comparisons and multi-comparisons. In both of cases the most important part of results is presented as a table of aligned sequences which are related with rows of the table. Columns of the table are related with "common positions" of the sequences: "to compensate differences in the sequence lengths" some positions in some sequences present "gaps" of amino acids. Distributions of gaps are organized in order that positions without gaps present "most similar" amino acids from different sequences [4, 13, 15]. In the first case two sequences are under consideration, and such procedures are formulated as optimization procedures which solutions are achieved by dynamic programming methods efficiently. Multi-alignment problems can be also formulated similarly as optimization tasks, but they cannot be solved efficiently [4,5,15,]. However, needs to have efficient multi-alignment procedures are so actual and important that practitioners have applied heuristic procedures which results can be estimated only indirectly and statistically by their testing on known data. There are several such procedures which are very popular

[1, 5, 27]. All of them use a pair-wise alignment as a preprocessing analysis which results after that are integrated in a multi-alignment result. The integration heuristic ideas are based on a statistical pattern about a profile related with a particular position in "correct, existed" (unknown) multi-alignment (in simplest case one can interpret the profile as an estimate of a probability distribution of amino acids for the position in evolution). The main difference between procedures is related with the profile heuristics.

We found only one exclusion from many different multi-alignment procedures, DIALIGN [26, 28, 29], which doesn't use any profile idea, and is based on absolutely different principle about a correct multi-alignment. It is a pure combinatorial principle in which the correct multi-alignment is combined as "a consistent" system of blocks. The whole system is estimated by a statistical objective function which extreme value is related with the correct alignment. It is interestingly that the function is designed without penalties for gaps. The gaps arise "automatically" from constrains on consistency of the system-alignment. Authors proposed an efficient greedy procedure to find a local maximum of the function (exact global solution is also NP-hard problem).

In this paper we analyze main combinatorial structures of DIALIGN and show another ways how they can be applied to construct new multi-alignment procedures. A novelty of these ways is they present multi-alignment results with some "inner structure" of the alignment, which can help in an interpretation analysis.

The paper is organized in 12 sections. In sections 2-6 we describe main DIALIGN's combinatorial elements [22, 30]. In section 7 we discuss a fundamental notions of layered clusters [23], on which our concept is based. In section 8 we introduce a quasi-concave measure of alignment quality. In section 9 we discuss a concept of the best multiple alignment choice. In section 10 the best choice principle is implemented algorithmically. In section 11 an illustrative computed example of multiple alignment is shown and discussed. The conclusions are given in section 12.

## 2. Alignment as an Equivalence Relation

Consider a set of sequences $\{X^{(1)},\dots,X^{(N)}\}$, where $X^{(i)}=<x^{(i)}_1,\dots,x^{(i)}_{Li}>$ is the $i$-th sequence and $Li$ is the length of $X^{(i)}$. Let us call $X^{(i)}_{int}=<x^{(i)}_k, x^{(i)}_{k+1},\dots,x^{(i)}_{k+p}>$ as *interval* of $X^{(i)}$.
Consider the set of pairs of indices

$$S=\{(i,\ l)\mid 1\le i\le N,\ 1\le l\le Li\}$$

of all the positions of all sequences (Sequence-Position Pair Set (*SPPS*)).

Consider a relation $R$ on $S$. Say that each two pairs $x,y\in S$ has a *connection* if $(x,y)\in R$.

Let $R$ be an equivalence relation on $S$. It means that the set S divided into $r$ classes $S_1,\dots,S_r$, $S_i\cap S_j=\varnothing$, $1\le i,j\le r$, $\bigcup_{i=1}^{r} S_i = S$ and two pairs $x_1=(i_1,l_1)$ and $x_2=(i_2,l_2)$ are in a relation $R$ (symbolized as $(x_1,x_2)\in R$) if and only if both of them belong to one of these classes.

Introduce a partial order "$\le$" relation on $S$ as following:

$$(i,l)\le(j,k)\Leftrightarrow i=j, l\le k,$$

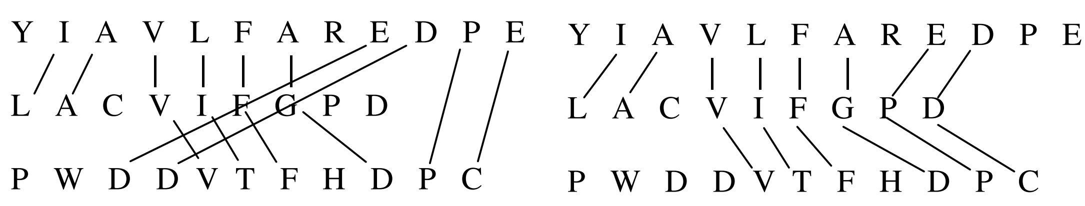
which means that only pairs from a same sequence can be compared.
Introduce a new relation on $S$:

$$\le_R := (R\cup\le)_t\,,$$

which is a transitive hull of union of $R$ and "$\le$" relations, i.e. the elements $c_1,c_2\in S$ are in relation $(R\cup\le)_t$ if and only if there exists a chain of elements $s_0,\dots,s_k\in S$ with $c_1=s_0$, $c_2=s_k$ and $(s_{i-1},s_i)\in R$ or $s_{i-1}\le s_i$, for all $i=1,\dots,k$.

An equivalence relation $R$ on the set $S$ is called *alignment* if all restrictions of the extended relation "$\le_R$" to the single sequences coincide with their respective natural order relation. In other words an equivalence relation $R$ on the set $S$ is called *alignment* if for any two pairs $x_1=(i,l_1)$ and $x_2=(i,l_2)$ from a same proper sequence $i$ the following is correct:

$$x_1\le x_2\Leftrightarrow x_1\le_R x_2.$$




(a) (b)

Figure 1: Two examples of an equivalence relation, which is an alignment (b) and an equivalence relation, which is not alignment (a).

Two examples of the introduced notions are given on the Figure 1. Here and below we numerate sequences from the top to the bottom. Lines reflect the connections on the set $S$. It is easy to see, that the restriction of "$\leq_R$" to the first sequence does not preserve its' natural order relation in the case (a). On the other side all the restrictions are compatible with the natural order relations of the single sequences in the case (b). These examples are important for understanding of all the construction thus let us discuss them in more details. Obvious that all the pairs connected with a line can be united into the equivalence classes. Each letter, which does not connect to any other one, forms its own class of equivalence. Let us take, for example, a path (1,10), (3, 4), (3, 8), (2, 6), (1,6) on the Figure 1 (a). We see that (1, 10) $\leq_R$ (1,6). But from the other hand (1, 10) > (1, 6) according to the natural order relation on the sequence 1. This observation shows us that an equivalence relation example illustrated by Figure 1 (a) is not an alignment.

Let us chose an arbitrary $x\in S$, an arbitrary $i$-th sequence, and an equivalence relation $R_A$, which is an alignment. The following boundaries characterize such triple:

$$b^{\downarrow}(i,x,R_A) := \min(j \in \{1,\ldots,Li+1\} \mid j = Li+1 \text{ or } x \leq_R (i,j)),$$
$$b^{\uparrow}(i,x,R_A) := \max(j \in \{0,\ldots,Li\} \mid j = 0 \text{ or } (i,j) \leq_R x).$$

Let us illustrate this characterization with help of Figure 1 (b). Let us consider sequences 1 and 3 and the letter R in the 1-st sequence. In this case $b^{\downarrow}(3, (1, 8), R_A) = 10$ and $b^{\uparrow}(3, (1, 8), R_A) = 9$. It means the letter R from the sequence 1 can be connected in $R_A$ with the sequence 3 only by a new position between original positions 9 and 10. In other words, only with a gap which should be inserted here.

Now introduce the notion of consistency of a pair of elements $x,y\in S$ to an alignment $R_A$.

A pair of elements $x,y\in S$ be *consistent* with an alignment $R_A$ if the relation $R_A' = R_A \cup \{(x,y)\}$ is also an alignment.

Describe now a way of checking that a pair of elements $x,y\in S$ is consistent with an already built alignment $A$. This checking procedure will be a part of the proposed multiple alignment algorithm.

Assertion [22]. Given an alignment $R_A$, an element $\hat{x}\in S$ and an element $\hat{y} = (i_1, j_1) \in S$, the relation $R_A' := R_A \cup \{(\hat{x},\hat{y})\}$ is a consistent extension of $R_A$ if and only if one has $b^{\uparrow}(i_1,\hat{x},R_A) < j_1 < b^{\downarrow}(i_1,\hat{x},R_A)$. Moreover, the consistency bounds of $R_A'$ can be computed, for any $x = (i_0, j_0) \in S$ and $i$, $1\leq i\leq N$ by the formulae

$$b^{\downarrow}(i,x,R_A') = \begin{cases} b^{\downarrow}(i,\hat{y},R_A) & \text{if} \quad b^{\uparrow}(i_0,(i,b^{\downarrow}(i,\hat{y})),R_A) < j_0 \leq b^{\uparrow}(i_0,\hat{x},R_A) \\ b^{\downarrow}(i,\hat{x},R_A) & \text{if} \quad b^{\uparrow}(i_0,(i,b^{\downarrow}(i,\hat{x})),R_A) < j_0 \leq b^{\uparrow}(i_0,\hat{y},R_A) \\ b^{\downarrow}(i,x,R_A) & \text{else.} \end{cases}$$

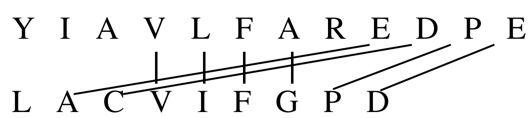
$$b^{\uparrow}(i,x,R_A') = \begin{cases} b^{\uparrow}(i,\hat{y},R_A) & \text{if} \quad b^{\downarrow}(i_0,(i,b^{\uparrow}(i,\hat{y})),R_A) > j_0 \geq b^{\downarrow}(i_0,\hat{x},R_A) \\ b^{\uparrow}(i,\hat{x},R_A) & \text{if} \quad b^{\downarrow}(i_0,(i,b^{\uparrow}(i,\hat{x})),R_A) > j_0 \geq b^{\downarrow}(i_0,\hat{y},R_A) \\ b^{\uparrow}(i,\hat{x},R_A) & \text{else.} \end{cases}$$

## 3. **Diagonals of Dot Matrices**

Now introduce a notion of *diagonal*. Given two sequences $X=<x_1,\ldots,x_{L1}>$ and $Y=<y_1,\ldots,y_{L2}>$ of length $L1$ and $L2$ respectively and $A=||a_{ij}||_{L1*L2}$ ,where $a_{ij}$ is a measure of compatibility of the $x_i$ and $y_j$. $A$ is called a dot matrix of $(X,Y)$.Each sequence of pairs $d=<(x_i,\ y_j),\ (x_{i+1},y_{j+1}),\ldots,(x_{i+k},y_{j+k})>,\ 1\le i\le L1,\ 1\le j\le L2,\ 0<k\le min(L1\text{-}i,\ L2\text{-}j)$ is called a *diagonal* of dot matrix.

The notion of diagonals is illustrated on the Figure 2. We present here the two first sequences from Figure 1, but with changed set of connections. Note that up to now we considered the notions of

| | L | A | C | V | I | F | G | P | D |
|---|---|---|---|---|---|---|---|---|---|
| Y | | | | | | | | | |
| I | | | | | | | | | |
| A | | | | | | | | | |
| V | | | | 1 | | | | | |
| L | | | | | 1 | | | | |
| F | | | | | | 1 | | | |
| A | | | | | | | 1 | | |
| R | | | | | | | | | |
| E | | 3 | | | | | | | |
| D | | | 3 | | | | | | |
| P | | | | | | | | 2 | |
| E | | | | | | | | | 2 |




(a) (b)

Figure 2: An Example of Diagonals in table (a) and graph (b) presentations. No alignment based on the 1-st and the 3-nd diagonals possible. An alignment based on the 1-st and the 2-nd diagonals is:

Y I A V L F A R E D P E
L A C V I F G - - - P D

An alignment based on the 2-nd and the 3-rd diagonals is:

Y I A V L F A R E D - - - - - - - - P E
L - - - - - - - - A C V I F G - - - P D

consistency and alignment in the common case of a group of sequences, and we called as alignment a multiple alignment of this group. On the Figure 2 we illustrate these notions in the case of only two sequences. One can see 3 diagonals on this figure. It is shown that it is impossible to build an alignment using 1-st and 3-rd diagonals. In opposite, the pair of diagonals 1 and 2, or the pair of diagonals 2 and 3 produce alignments.

Let us consider a relation $Q$ on a SPPS from the set $S$, such that

(a) $(x,\ y)\in Q \Rightarrow (y,\ x)\in Q,$

(b) $(x,\ y)\in Q,\ (x,\ z)\in Q \Rightarrow y = z,$

(c) $(y,\ x)\in Q,\ (z,\ x)\in Q \Rightarrow y = z,$

and a pair of sequences $X^{(i)}$ and $X^{(j)}$. Choose any interval from a sequence $X^{(i)}$ . Let $X^{(i)}{}_{int}$ be an interval from the sequence $X^{(i)}$ without gaps and $X^{(j)}$ is some another sequence which is under consideration. It is easy to show that if every pair $(i,\ l)\in X^{(i)}{}_{int}$ is associated with such pair $(j,\ k)$ from $X^{(j)}$ that $((i,\ l),\ (j,\ k))\in Q$ then

$X^{(i)}_{int}$ is associated with an interval $X^{(j)}_{int}$ in order that ($X^{(i)}_{int}$, $X^{(j)}_{int}$) is a diagonal. In opposite, each diagonal induces connections on the set $S$. The set of these connections is a relation $Q$ with *(a), (b), (c)* properties.

One can examine 6 diagonals on Figure 1 (a) and 5 diagonals on Figure 1 (b).

Now it is easy to see that an alignment consists of a set of diagonals. A set of diagonals is called an *alignment* if the equivalence relation influenced by it is an alignment.

Checking if a subset of diagonals $G=\{g_1,\ldots,g_k\}$ is an alignment can be done by the following recursive procedure:

*Step 0:* $i=0$, $A=\varnothing$;
*Step i:* $i=i+1$, $A=A\cup g_i$, if all the positions of $g_i$ are successively checked out for the consistency with the widened set of positions. If one found a non-consistent position, then $G$ is non-consistent. Stop.
If $i=k+1$ then $G$ is consistent. Stop.

Using a set of diagonals which is an alignment, one can easily built a multiple alignment matrix, with matching induced by the diagonals and gaps or letters, which are not included into the diagonals between these intervals of matching. Let us illustrate it on the alignment presented on the Figure 1 (b).

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Y** | **I** | **A** | **-** | **V** | **L** | **F** | **-** | **A** | **R** | **E** | **D** | **P** | **E** |
| **-** | **L** | **A** | **C** | **V** | **I** | **F** | **-** | **G** | **-** | **P** | **D** | **-** | **-** |
| **P** | **W** | **D** | **D** | **V** | **T** | **F** | **H** | **D** | **-** | **P** | **C** | **-** | **-** |

Figure 3: A multiple alignment corresponded to the set of diagonals presented on Figure 1 (b).

Note that any successive widening of alignment pair by pair according to the Assertion [22] while the alignment constraints are succeed gives a multiple alignment.

## 4. Diagonal Weighting

Consider two ways of a diagonal weight calculation. Let *D* is the set of all the considered diagonals and *d* is an arbitrary diagonal from the set *D*.

In case that one use a 0-1 dot matrix *w(d)* can be calculated as following. Let *l* is the length of *d*, m is number of matches in *d*, *T* is a parameter.

$$w(d) = \begin{cases} E(l,m), \textit{if } E(l,m) > T, \\ 0, \text{otherwise} \end{cases},$$

$$E(l,m) = -\ln P(l,m),$$

$$P(l,m) = \sum_{i=m}^{l} \binom{l}{i} p^i (1-p)^{l-i}.$$

*P(l,m)* is the probability of diagonal of length *l* has equal to or greater than *m* matches. *p*=0.05 in the protein case and *p*=0.25 in the DNA case.

If one use a substitution dot matrix then *w(d)* can be calculated as:

$$w(d) = \begin{cases} E(S), \textit{if } E(S) > T, \\ 0, \text{otherwise} \end{cases}$$

$$E(S) = -\ln P(S)$$

$$P(S) = 1 - e^{-K*N^2 e^{-\lambda S}},$$

$$S = \max_{d_{sub} \subseteq d} \sum_{(i,j) \in d_{sub}} a_{ij},$$

$$N = max(l- \quad \lambda S/H,\ 1).$$

*P(S)* is the probability of a length *l* diagonal maximal hit score is greater than or equal to *S* (extreme value distribution). *K, H,* $\lambda$ are functions of dot matrix elements $a_{ij}$ and a prior probability distribution for amino acids [24].

## 5. Pseudo Diagonals

An Alignment Quality Measure is a function of its individual diagonal weights and the weights of "diagonal interactions." In order to take these interactions in consideration the following construction of pseudo diagonal is introduced.

Let *w(d)* be a weight of diagonal *d,* and *e* is another diagonal.

Definition. If *d* and *e* have a common sequence S1, then there are 2 sequences (S1,S2) for *d* and (S1,S3) for *e*. Mark all the positions of S1 that included in *d* and *e* simultaneously. We call the new diagonal, which contains positions of S2 and S3 connected to the marked ones of S1 as *pseudo diagonal z*. Thus the interaction between *d* and *e* can be measured as following:

$$\tilde{w}(d,e) = \begin{cases} w(z) & \textit{if } z \text{ exists}, \\ 0 & \text{else}. \end{cases}$$

## 6. The Set-Theoretical Description of Consistent Alignments

Let us consider the set of diagonals $D$ and the set $C$ of all the alignments on $D$: $C \subseteq 2^D$. The following statements can be easily checked:

1. $\varnothing \in C$;
2. $C$ is a lower semilattice of sets: $H_1, H_2 \in C \Rightarrow (H_1 \cap H_2) \in C$;
3. $H_1 \in C$, $H_2 \subseteq H_1 \Rightarrow H_2 \in C$. Let $\{E_1, \ldots, E_m\}$ be the set of maximal elements of $C$. Then $C$ is a union of the set-theoretical intervals:

$$C = \bigcup_{i=1}^{m} [\varnothing, E_i].$$

## 7. Layered Clusters

According to Mullat [25] and Mirkin, Muchnik [23] the tightness function, $F(H)$ on a subset $H$ of a finite set $W$, can be defined as following:

$$F(H) = \min_{i \in H} \pi(i, H),$$

where $\pi(i, H)$ relates any non-empty subset $H \subseteq W$ with its element $i \in H$. If $\pi(i, H)$ is a monotone linkage function: $H_1 \subseteq H_2 \Rightarrow \pi(i, H_1) \leq \pi(i, H_2)$ then the corresponding tightness function satisfies the so-called *quasi-concavity* condition: for any H,G∈ W,

$$F(H \cup G) \geq \min(F(H), F(G)).$$

In this work we are interested only in the class of monotone linkage functions and the corresponding quasi-concave tightness functions.

The tightness function reflects the density of interrelations within sets $H$.

Let us refer to a subset $H \subseteq W$ as to a *pattern* with regard to $F(H)$ if $H$ is separated from the rest in such a way that $F(H)$ is greater than $F(H')$ for any $H'$, which is not its part, that is, $F(H) > F(H')$ for any $H \subseteq W'$ such that $H' \cap (W-H) \neq \varnothing$.

The set of all patterns $P_{all}$ is nonempty and chain-nested: $W=P_1 \supset P_2 \supset \ldots \supset P_k=H^*$. $H^*$ here is the maximal global maximizer of $F(H)$: $F(H^*) \geq F(H)$, for every $H \subseteq X$ and $F(H^*) > F(H)$, for every $H \not\subseteq H^*$. $H^*$ named *kernel* of quasi-concave function.

$P_{all}$ is named layered cluster. The pattern of a layered cluster can be considered as levels of resolution of the overall similarity modeled by the set function F.

The following worst-out *Algorithm 1*:

1. $X_{curr}=X$;
2. $x = arg\ min\ \pi(x, X_{curr}), x \in X_{curr}$;
3. $X_{curr}=X_{curr} \backslash \{x\}$;
4. If $X_{curr}= \varnothing$, then STOP, else go to 2,

successively built $P_{all}$: $X_{curr}$ is a pattern iff $F(X_{curr}) > \max\limits_{Y = X_{curr} \text{ on the previous steps}} F(Y)$.

It is easy to check that all the theory can be transferred to the set theoretic intervals $[A, W]$ by redefinition $\pi'(i, H) = \pi(i, H \cup A)$.

## 8. An Alignment Density Measure

Let $H$ is a subset of $D$: $H \in 2^D$. Let d is a diagonal from $H$: $d \in H$. Introduce the following linkage function:

$$\varphi(d, H) = w(d) + \sum_{e \in H} \tilde{w}(d, e),$$

which scores a linkage between a subset $H$ and a diagonal $d$. It is easy to see that $\varphi(i, H)$ is a monotone linkage function: $H_1 \subseteq H_2 \Rightarrow \varphi(i, H_1) \leq \varphi(i, H_2)$.

Introduce the tightness function

$$\Phi(H) = \min_{i \in H} \varphi(i, H),$$

which is quasi-concave correspondingly. This function reflects the density of interrelations within sets $H$ in such a way that the greater $\Phi(H)$ the greater the density of $H$.

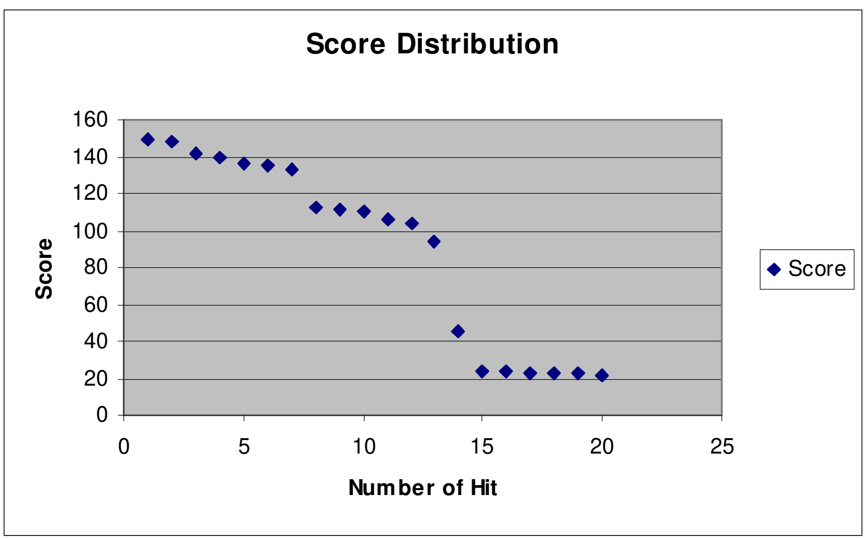



Figure 4: Hit scores distribution in a database search. X axe reflects the order of hits in a database search output. The hit score values marked on the Y axe. The clustering of hits is observed.

## 9. The Multiple Alignment Choice: Concept and Formalization

Let us start to expose the concept from the following simple observation. Each one that made a protein sequences database search at least once is familiar with it. Given a query sequence and a database and one do the homologous search using a local alignment algorithm. The result of search is a chain of hits from the database sequences sorted according the descending of the scores. One can observe that the significant hits are clustered in the following way: the most closed homologues formed the core, the first cluster in the scores order then one can observe a shift on the score axe. After it a cluster of hits, which closeness to the query is weaker is observed. This cluster is also ended by shift in score, and so on.

We illustrate the exposed observation on Figure 4.

This observation suggests a nested structure of significant hit sets in a group of homologues. The closest hits will be clustered together, the second significant group of hits will be clustered together and with the first group and so on.

Let us illustrate this consideration with help of the following example. On Figure 5 one can see three sequences and a imaginary results of pairwise alignments of the sequences from Figure 1 (a), which produce the pairwise diagonals. If one unite all the diagonals on a same picture (Figure 5 (b) which is the same as Figure 1 (a)) then one can see that this set of diagonals is incompatible and can not produce an alignment. It can be easily shown that only two set-theoretical intervals of alignments can be extracted:

$$I_1 = \left\{ \varnothing, \quad \left\{ \begin{pmatrix} [1,2] \\ \text{IA} \\ \text{LA} \end{pmatrix}, \begin{pmatrix} [1,2] \\ \text{VLFA} \\ \text{VIFG} \end{pmatrix}, \begin{pmatrix} [1,2] \\ \text{ED} \\ \text{PD} \end{pmatrix}, \begin{pmatrix} [2,3] \\ \text{VIF} \\ \text{VTF} \end{pmatrix}, \begin{pmatrix} [2,3] \\ \text{GPD} \\ \text{DPC} \end{pmatrix} \right\} \right\}, \text{ and}$$

$$I_2 = \left\{ \varnothing, \quad \left\{ \begin{pmatrix} [1,2] \\ \text{IA} \\ \text{LA} \end{pmatrix}, \begin{pmatrix} [1,3] \\ \text{ED} \\ \text{DD} \end{pmatrix}, \begin{pmatrix} [1,3] \\ \text{PE} \\ \text{PC} \end{pmatrix}, \begin{pmatrix} [2,3] \\ \text{GPD} \\ \text{DPC} \end{pmatrix} \right\} \right\}.$$

Let us analyze the multiple alignment matrices, which correspond to the maximal sets of these two intervals. These matrices are presented on Figure 6. The first alignment looks better than the second one, because it has more matching and much less gaps. What another difference we see between these two pictures? On the first picture we see a dense cluster in the middle of it, the two rather dense segments which are connected to the first cluster, the segment connected to the second pair and so on.

The second alignment consists of three different blocks, which are not connected together, that's why this alignment is a weak one.

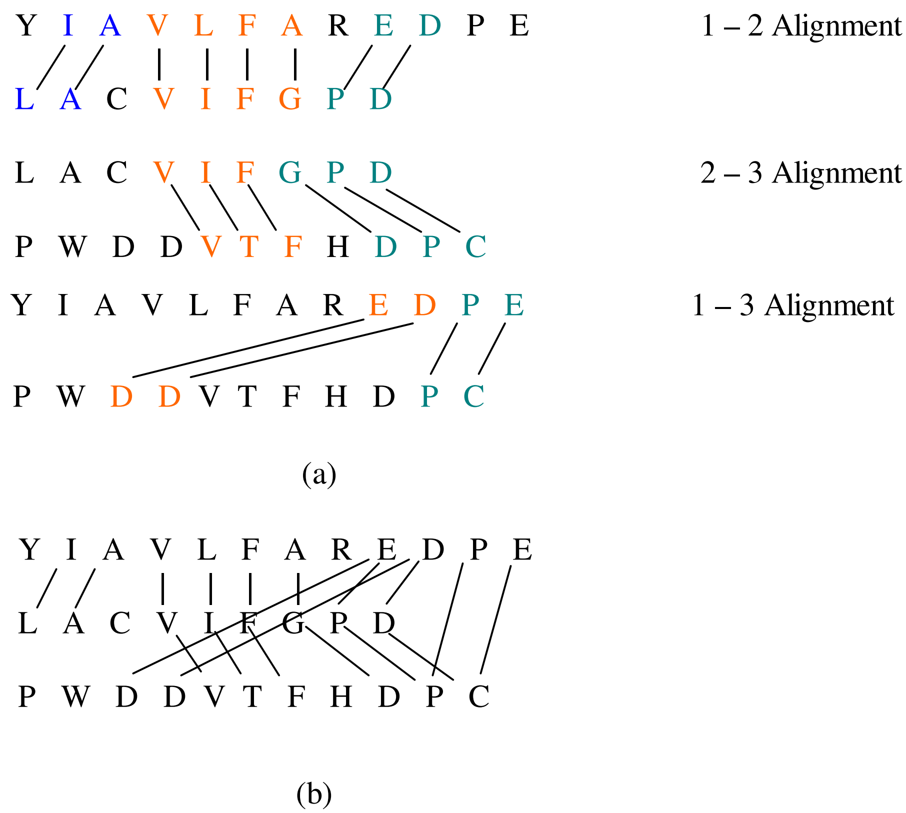


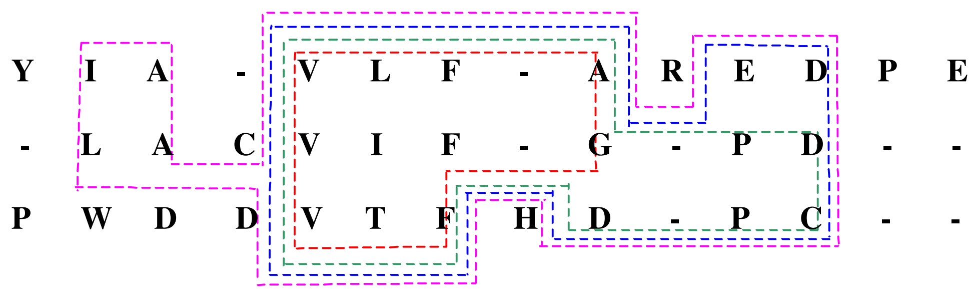

Figure 5: (a) three pairwise alignments separately. Diagonals are colored in the same color and their corresponding amino acids are connected with a line. (b) All the diagonals obtained in three alignments united on the one picture.

According to the enclosed structure of hits, illustrated above, we consider a multiple sequence alignment as a crystallization process: its first step is the nucleus of crystal formation and the further process is a crystal growing layer by layer. The nucleus of group of protein sequences is the set of the segments of its individual sequences with large connection strength, the connection strength decreasing in order of the layers' addition.

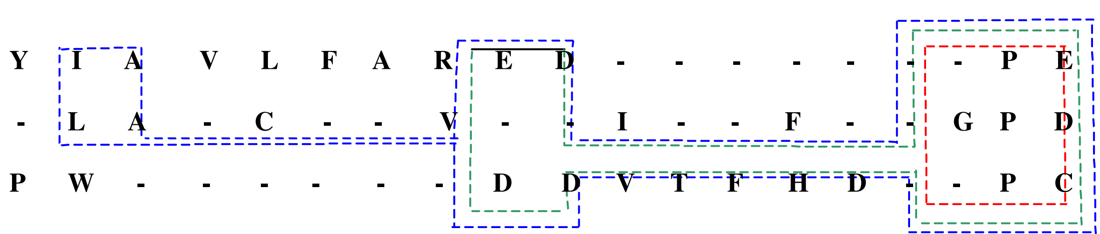
In other words, a procedure which restores this process has to find firstly the most dense set of hits, than to add the less dense, but mostly connected with the first set and so on.

This process can be modeled as a restoration of such a set theoretical interval $[\varnothing, E^*]$ which contains a chain of local extremes of the function $\boldsymbol{\Phi}$:

$$
\begin{aligned}
&P_p = \arg\max \Phi(H \mid H \in C);\\
&P_{p-1} = \arg\max \Phi(H \mid P_p \subseteq H \in C);\\
&\ldots\\
&P_0 = E^*.
\end{aligned}
$$




(a)




(b)

Figure 6: (a) an alignment corresponded to the maximal element of the first set-theoretical interval. (b) an alignment corresponded to the maximal element of the second set-theoretical interval.

## 10. An Algorithm of the Crystallized Multiple Sequence Alignment Reconstruction

We propose the following heuristics using the worst-off greedy *Algorithm 1* of patterns search on a set theoretic interval:

*Algorithm 2:*

$E=\varnothing$, $H^*=\varnothing$, $H^*_{prev}=\varnothing$;

*Forward:*

do

$E=E\cup\{i\}$ where $i = arg\ max\ \Phi(E\cup\{i\})$ over $i\in D\backslash E$, $i$ is consistent with $E$

while such i exists;

*Backward:* $H^*_{prev}=H^*$. Find $H^*=arg\ max\ \Phi(H)$ over $[\varnothing, E]$ with help of the *Algorithm 1*. If $H^*=E$, then stop and output $E$ and all the chain of $H^*$. If $H^*_{prev}\neq H^*$ then $E=H^*$ and go to *Forward.* If $H^*_{prev}=H^*$ then find $H^*=arg\ max\ \Phi(H)$ over $(H^*, E]$ with help of the *Algorithm 1*. If $H^*=E$, then stop and output $E$ and all the chain of $H^*$. Else $E=H^*$ and go to *Forward.*

Using less formal words, this algorithm works in the following way: starting from the empty set it runs forward providing a greedy choice of the maximal value of $\Phi$ on each step. It corresponds to an intuition for choice of an interval with big values of $\Phi$ on its successive nested sets. When it reaches an upper consistent set it turns back and found the set maximizes $\Phi$ on the chosen interval. Starting from this maximizer it repeats the forward and the backward passes until two successive passes give the same

```
HTLV_II  t PLPSH e t hSAQKGELLALI CGL r aAKPWPS- -LNIFL–  DSKYLIK--- - YLH sl  a i
RSV       lGA- -- ----SVQQLEARAVAMAL L– L- - WPTTPTNVVT– DSAFV a k mll  k m - - -- GQ
MoMLV    AGT- - - ----SAQRAELIALTQALK-M- - AEGKKLNVYT– DSRYAFA  - -- - TAH ih GE
HBV       a PLPIH- - -T- - - AELLA a- C f - - -ARSRSG- -ANI IGTDN- ----  -- - - -- --- -
E.coli   AGYT r T- - -TNNRMELMA AI V ALE a LK- - E h c e v i l s T– DSQY vr q  g itQ WIH ---
```

**[2, 3]**
**GASVQQLEARAVAMALLLWPTTPTNVVTDSAFV**
**GTSAQRAELIALTQALKMAEGKKLNVYTDSRYA**
**[1, 3]**
**LNIFLDSKYLIKYLH**
**LNVYTDSRYAFATAH**
**[3, 5]**
**TDSRY**
**TDSQY**

[3, 4] *TDS* *TDN*

[1, 2] *SAQKGELLALICGL* *SVQQLEARAVAMAL*

[3, 5] *TSAQRAELIALTQALK* *TTNNRMELMAAIVALE*

[3, 4] *AELIA* *AELLA*

[1, 4] *PLPSH* *PLPIH*

[1, 4] *AKPWPSLNIF* *ARSRSGANII*

[1, 2] WP WP

[3, 5] AG AG

[1, 4] C C

[2,3] GQ GE

[3, 5] MAE LKE

Figure 7: Multiple alignment results. The nucleus picked out by bold, the layers with medium density italicized and the layers with small density printed with regular font.

solution. When this search is finished the nucleus is fixed and the following layer is found in the same way. The procedure stops when the last layer, which is a maximal consistent set is found.

## 11. Illustrative Results

In order to test the proposed concept we implemented the Algorithm 2 in software and processed a group of 5 protein sequences. The first step of our processing was all 10 pairwise alignments built with DIALIGN algorithm for each pair of sequences. The result of this step was a set of all possible diagonals, which was used for further processing. On the second step we ran the Algorithm 2 implementation on this set. On the last step our software places the obtained set of diagonals on the alignment matrix, with gaps signed as "-". The results of the processing presented on Figure 7. We obtained 11 layers and classified them into 3 groups: nucleus, medium density, and small density. Figure 7 (a) illustrates the "forms" of layers. One can see that the "forms" rather complicated, with "nonlinear boundaries." It corresponds to our concept of the crystallize growth.

In order to estimate how distant our results from the results of the known algorithms we compared them with the ClustalW [27] results and with the DIALIGN results. One can visually compare these results presented on Figure 8. To make this comparison more formal we compare multiple alignments according to the following criteria:

1. Choose all the columns which have >1 amino acid. Count the number of columns, which have only 2 amino acids and these amino acids are identical. Count the number of columns, which have more then 2 amino acids and ≥ 2/3 amino acids in a column are identical. Summarize these two counters.
2. Choose all the columns which have >1 amino acid. Count the number of columns, which have only 2 amino acids and these amino acids or both hydrophobic or both hydrophilic. Count the number of columns, which have more then 2 amino acids and ≥ 2/3 amino acids in a column are or all hydrophobic or all hydrophilic. Summarize these two counters.
3. Count the number of gaps in a multiple alignment.

| Algorithm | Criteria 1 | Criteria 2 | Criteria 3 |
|---|---|---|---|
| Our algorithm | 20 | 21 | 21 |
| ClustalW | 12 | 19 | 9 |
| DIALIGN | 20 | 19 | 15 |

Table 1: Comparison results

The results of the comparison summarized in the Table 1. One can see that our results are very close to the DIALIGN results, but a little worse in the 3-rd criteria. The DIALIGN uses the same principle of diagonal set consistency. ClustalW does not provide good matching because of it is a global one. We lose a little for DIALIGN, because of its iterative structure. In terms of iterations we implemented only the first one.

We conclude that the proposed principle gives rather good results and it is expedient to continue research of its properties.

## 12. Conclusion

We had two motivations to propose the above multi-alignment procedure:

a. to emphasize that "correct" multi-alignment should be an analog for a crystallized process based on finding "a layered cluster" structure on the aligned set of sequences because such structure is reflected in "a nature" of data for what one wants to get a multi-alignment representation (according to the motivation we could show both that it is possible and that it gives a new way to construct multi-alignments);

```
HTLV_II  TPLPSHETHSAQKGELLALICGLRAAKPWPS- - LNIFL–DSKYLIK- - - -YLHSLAI
RSV      LGA- - - - - -SVQQLEARAVAMALL–L- -WPTTPTNVVT–DSAFVAKMLLKM- - - -GQ
MoMLV    AGT- - - - - -SAQRAELIALTQALK-M- -AEGKKLNVYT-DSRYAFA- - - -TAHIHGE
HBV      APLPIH- - -T- - -AELLAA-CF- - -ARSRSG- -ANIIGTDN- - - - - - - - - - - - - - - -
E.coli   AGYTRT- - -TNNRMELMAAIVALEALK- -EHCEVILST-DSQYVRQGITQWIH- - - -
```

(a)

```
HTLV_II  TPLPSHETHSAQKGELLALICGLRAAKPWPSLNIFLDSKYLIKYLHSLAI
RSV      - - - - - -LGASVQQLEARAVAMALLLWP-TTPTNVVTDSAFVAKMLLKMGQ
MoMLV    - - - - - -AGTSAQRAELIALTQALKMAE-GKKLNVYTDSRYAFATAHIHGE
HBV      - - - - - -APLPIHTAELLA- -ACFARSRSGANI-IGTDN- - - - - - - - - - - - - -
E.coli   - - -AGYTRTTNNRMELMAAIVALEALKEHCEVILSTDSOYVROGITOWIH
```

(b)

```
HTLV_II  TPLPSHETHSAQKGELLALICGLRAAK- -PWPSLNIFLDSKYLIK- -YLHSLAI
RSV      LG- - - - - - -ASVQQLEARAVAMALLLWPTTP- - -TNVVTDSAFVAK- -MLLKMGQ
MoMLV    AG- - - - - - -TSAQRAELIALTQALKMAEGKK- - -LNVYTDSRYAFA- -TAHIHGE
HBV      APLPIHT- - - - - -AELLAACFA-RSRSGAN- - - -IIGTDN- - - - - - - - - - - - - -
E.coli   AG- - -YTRTTNNRMELMAAIVALEALK- -EHCEVILSTDSQYVRQGITQWIH- -
```

(c)

Figure 8: (a) our results. (b) ClustalW results. (c) DIALIGN results. Columns with ≥ 2/3 equal amino acids are bolded. Columns with ≥ 2/3 hydrophobic or ≥ 2/3 hydrophilic amino acids are italicized.

b. the combinatorial optimization technique for quasi-concave set functions an appropriate, useful and efficient approach in order that to be applied in a sequence multi-alignment analysis.

Experimental illustration has showed that this is a promised approach. Moreover, the procedure can be differently modified and improved using the same basic ideas. For instance, instead the set of diagonals one can consider all possible sub-diagonals from all given diagonals. Another example, instead using just a unique FORWARD, one can adapt the procedure FORWARD for starting from all possible diagonals, and, keep as a final result the best according to the maximum value of the function Φ(H). We want to investigate such possibilities in detail.

The main part of our plan on the near future is to extend seriously data for the approach testing and comparative analysis.

## REFRENCES


1. Carrillo & Lipman, (1988) The Multiple Sequence Alignment Problem in Biology. SIAM J. Appl. Math. 48 1073-1082;
2. Altschul & Lipman, (1989) Trees, Stars, and Multiple Biological Sequence Alignment. SIAM J. Appl. Math. 49, 197-209
3. Altschul (1989) Gap Costs for Multiple Sequence Alignment. J. Theor. Biol. 138, 297-309
4. Cuff, J. A. and Barton, G. J. (2000), Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction", PROTEINS: Structure Function and Genetics, 40:502-511.

5. Barton,G. J. (1998) Protein sequence alignment techniques,
Acta Cryst.D.,54,1139-1146

6. Bowie JU, R, Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure
Science 253:164-170.

7. Greer, J. (1991),Comparative modeling of homologous proteins
Meth. Enzymol., 202, 239-252.

8. Henikoff, S. & Henikoff, J. G. (1992)Amino acid substitution matrices from protein blocks.
Proc. Natl. Acad. Sc. U.S.A., 89, 10915-10919.

9. Vingron, M. & Waterman, M. S. (1994) Sequence alignment and penalty choice. J. Mol. Biol., 235, 1-12.

10. Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence-structure threading. Curr. Opin. Str. Biol., 5, 236-244.

11. Rost WWW, B. (1996c). Protein fold recognition by merging 1D structure and sequence alignments. EMBL Heidelberg, Germany, WWW document
(http://dodo.bioc.columbia.edu/~rost/Papers/96PreTopits.html).

12. Fischer & Eisenberg(1996) Fold Recognition Using Sequence-Derived Predictions,Protein Science, 5, 947-955, 1996

13. Grundy, William N., Timothy L. Bailey, Charles P. Elkan and Michael E. Baker (1997) Meta-MEME: Motif-based Hidden Markov Models of Biological Sequences Computer Applications in the Biosciences, 13(4):397-406, 1997.

14. Yuan, Y.P., Eulenstein, O., Vingron, M. & Bork, P.(1998) Towards detection of orthologues in sequence databases. Bioinformatics, 14(3):285-9

15. Barton,G. J. (1998) Protein sequence alignment techniques, Acta Cryst.D.,54,1139-1146

16. Bork, P. & Koonin, E.V. (1999) Predicting functions from protein sequences--where are the bottlenecks? Nat Genet 1998 Apr 18(4) 313-318

17. Andrade, M.A. (1999) Position-specific annotation of protein function based on multiple homologs. ISMB 1999 7 28-33

18. Prlic,A., Domingues,F.S., and Sippl,M.J , (2000)
Structure derived substitution matrices for alignment of distantly related sequences. Protein Engineering, 13(8): 545-550 (2000)

19. Domingues, F.S., Lackner, P., Andreeva,A., and Sippl, M.J.
Structure based evaluation of sequence comparison and fold
recognition alignment accuracy. J. Mol. Biol., 297(4): 1003-1013 (2000)

20. Cuff, J. A. and Barton, G. J. (2000), Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction", PROTEINS: Structure Function and Genetics, 40:502-511.

21. Thompson J.D., Plewniak F., Thierry J.-C. and Poch O.( 2000) `DbClustal:` Rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acid Research, 2000, Vol.28, No 15 2919-2926

22. B. Morgenstern, J. Stoye, A. Dress (1999). Consistent equivalence relations: a set-theoretical framework for multiple sequence alignment. University of Bielefeld, FSPM, Materialien/Preprints 133.

23. B. Mirkin, I. Muchnik (2001, to appear) Layered Clusters of Tightness Set Functions and Their Relation to Hereditary Mappings and Convex Geometries.

24. S. Karlin, S.F. Altshul. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 90, 2264-2268

25. J. Mullat. (1976) Extremal subsystems of monotone systems: I,II. Automation and Remote Control, 37:758-766, 37:1286-1294.

26. B. Morgenstern, A. Dress, T. Werner. (1996).Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. USA 93, 12098-12103

27. J. D. Thompson, D.G. Higgins, and T.J. Gibson (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22:4673-4680.

28. B. Morgenstern, K. Frech, A. Dress, T. Werner (1998). DIALIGN: Finding local similarities by multiple sequence alignment. Bioinformatics 14, 290 - 294.

29. B. Morgenstern (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15, 211 - 218.

30. S. Abdedda(1997). Incremental Computation of Transitive Closure and Greedy Alignment. Proceedings of the 8-th Annual Symposium on Combinatorial Pattern Matching. Lecture Notes in Computer Science 1264, Springer Verlag, pp. 167 - 179.