A Procedure for Biological Sensitive Pattern Matching in Protein Sequences

Juan Méndez, Antonio Falcón, and Javier Lorenzo

Intelligent Systems Institute. IUSIANI
Univ. Las Palmas de Gran Canaria, Spain
[jmendez,afalcon,jlorenzo]@dis.ulpgc.es

Abstract. A Procedure for fast pattern matching in protein sequences is presented. It uses a biological metric, based on the substitution matrices as PAM or BLOSUM, to compute the matching. Biological sensitive pattern matching does pattern detection according to the available empirical data about similarity and affinity relations between amino acids in protein sequences. Sequence alignments is a string matching procedure used in Genomic; it includes insert/delete operators and dynamic programming techniques; it provides more sophisticate results that other pattern matching procedures but with higher computational cost. Heuristic procedures for local alignments as FASTA or BLAST are used to reduce this cost. They are based on some successive tasks; the first one uses a pattern matching procedure with very short sequences, also named k-tuples. This paper shows how using the L_1 metric this matching task can be efficiently computed by using SIMD instructions. To design this procedure, a table that maps the substitution matrices is needed. This table defines a representation of each amino acid residue in a n-dimensional space of lower dimensionality as possible; this is accomplished by using techniques of Multidimensional Scaling used in Pattern Recognition and Machine Learning for dimensionality reduction. Based on the experimental tests, the proposed procedure provides a favorable ration of cost vs matching quality.

Keyword: Pattern Matching, Biological Pattern Analysis, Sequence Alignments, Multidimensional Scaling, SIMD Processing.

1 Introduction

The fast growing of information contained in the biological databases[1] requires more efficient processing systems to find functionality and meaning in the DNA and protein sequences. More efficient systems are obtained by hardware and architectural improvements, and also by defining more efficient computational procedures. Artificial Intelligence, Pattern Recognition and Machine Learning techniques can provide additional approaches to allow better computational performances in Gemomic related systems[2]. This paper uses Pattern Recognition

and Machine Learning techniques applied in Bioinformatics[3] to define a matching procedure to get some architectural improvements in alignment procedures of biological sequences. These architectural improvements are initially introduced for multimedia and information retrieval applications, but by means of special software design they can also be used in genomic related computations.

Single Instruction Multiple Data(SIMD) instructions are included in most microprocessors of low cost computer systems, as Intel and AMD. They can be used to speed up workstations and servers in Genomic, but special designs are needed because available compilers do not take advantage of these instructions for general software. Modern computer items as cache hierarchy, memory access and SIMD processing upgrade the performance of generic software, but additional increase of the power in genomic based procedures can be obtained if they are designed according to the above processor characteristics[4]. Some works have dealt with the use of parallel computation for sequence analysis[5, 6], and also with the use of SIMD instructions in the improvements of local alignments[7, 8]. However, this work presents a process for the first stages of some local alignment procedures. The proposal requires the computation of some tables to map the amino acid residues in a n-dimensional space according to the biological properties represented in the score or substitution matrices, as PAM[9] and BLOSUM[10].

The search of local alignment between biological sequences is one of the most used tools in discovering the functional and evolutionary similarities. The Smith-Waterman procedure[11], based on dynamic programming, has the highest biological significance. However, its computational cost is greater than other heuristics procedures as FASTA[12] and BLAST[13] which have lower computational cost having a high level of biological significance. The first stage of both FASTA and BLAST is the searching of very short pre-coded sequences, named k-tuples, in the sequences included in the biological databases. The matching of k-tuples, named ktup in FASTA and w-mers in BLAST, between a query sequence and the database can be efficiently computed by information retrieval procedures.

However instead of naive ASCII code matching, a n-dimensional code matching based on the biological information contained in the score or substitution matrices is proposed in this paper. The information retrieval procedure takes advantage of two architectural improvements of modern microprocessors: parallel computation with multiple data processing units, and sequential memory access which increases the cache throughput. This paper present the process to map the amino acid residues in a virtual meaning less n-dimensional space. This is accomplished by non-linear dimensionality reduction methods used in Multidimensional Scaling(MDS)[14–18] which are mainly used in Pattern Recognition and Machine Learning for feature selection and also for visualization of high dimensional data sets.

2 Pattern Matching of k-tuples

An efficient procedure for pattern matching of k-tuples is proposed. The distance $D(U,V)$ between two vector U and V in $\mathbf{R}^M$ based on the L_1 norm is defined as:

$$D(U,V) = \sum_{i=1}^{M} \mid U_i - V_i \mid \tag{1}$$

The Intel IA-32 computer architecture includes an instruction to compute this distance with $M = 8$ in a single system clock cycle. The norm for $M = 8 \times m$ also can be fast computed from the previous. The continuous increasing of microprocessor clock frequency provides a powerful method to speed up many of data processing tasks which can be re-formulated to fit in a L_1 norm. This instruction is part of the MMX instruction set included to improve the performance of multimedia, text retrieval and signal processing applications. Most of problems related with sequence analysis are based on score matrices to model the amino acid distances and similarities; this is not an efficient choice to use the power that current hardware provides. If $\mathcal{A}$ is the amino acid symbols set, instead of using a score matrix $s(a,b); a,b \in \mathcal{A}$, a distance based on norm L_1 can be required:

$$D_X(a,b) = \sum_{i=1}^{n} \mid X_i(a) - X_i(b) \mid \tag{2}$$

where $\mathbf{X}(a)$ is a n-dimensional vector which is the representation of the amino acid, and $D_X(a,b)$ is the desired distance. In raw text searching of query sequence in a biological database, this vector is the 1-dimensional ASCII code of the residue symbol. However, this is a too simplistic representation of the amino acid properties which ignores the biological meaning and the affinity relations. The similarity relations of amino acid require the introduction of a representation in a multidimensional space with the lowest dimensionality as possible. This representation must contain the biological information of similarity which is gathered in the substitution matrices. PAM and BLOSUM matrices are defined from statistical properties related with residues substitutions from evolutionary or blocks alignments. They are nor distance neither similarity functions. They are score factors which verifies: $s(a,b) = s(b,a)$ and also generally: $s(a,a) \geq s(a,b)$. From a score matrix several distance functions, $d(a,b)$, can be proposed; the considered in this paper is:

$$d(a,b) = s(a,a) + s(b,b) - 2s(a,b) \tag{3}$$

This verifies the symmetrical property: $d(a,b) = d(b,a)$, is lower bounded: $d(a,b) \geq 0$ and also verifies: $d(a,a) = 0$, but is not a metric. When is verified that $s(a,a) > s(a,b)$, it is also verified that if $d(a,b) = 0$ it must be: $a \equiv b$. The triangular properties is not verified in the general case, thus the proposed function is a distance, but not a metric one. This distance has also a probabilistic

expression when is computed from the PAM and BLOSUM substitution matrices. Both are obtained by means of a probabilistic ratio obtained from different empirical environments. In these cases, the score matrix and the distance are defined as:

$$s(a,b) = \frac{1}{\lambda}\log\frac{p(a,b)}{p_a p_b} \qquad d(a,b) = -\frac{2}{\lambda}\log\frac{p(a,b)}{\sqrt{p(a,a)p(b,b)}} \tag{4}$$

where $p(a,b)$ is the probability of substitution between two residues, p_a term is defined from the $p(a,b)$, and λ is a suitable parameter. The score of a k-tuple of two sequences U and V is computed in the alignment procedures[11, 19] by using substitution matrices as:

$$s(U,V) = \sum_{j=1}^{k} s(u_j, v_j) \tag{5}$$

where $u(j)$ and $v(j)$ correspond to the amino acid in the k-tuple. If the distance of this k-tuple, $d(U,V)$, is defined as: $d(U,V) = s(U,U) + s(V,V) - 2s(U,V)$, it can be computed as:

$$d(U,V) = \sum_{j=1}^{k} d(u_j, v_j) \simeq D_X(U,V) = \sum_{j=1}^{k}\sum_{i=1}^{n} \mid X_i(u_j) - X_i(v_j) \mid \tag{6}$$

If $d(a,b)$ can be computed by $D_X(a,b)$ with a reduced error. This last is a L_1 norm with $M = n \times k$. Due to hardware constraints, the optimal computation can be achieved when $n \times k = 8 \times m$. The high k value reduces the sensibility whereas the low k value implies a lower significative; BLAST uses $k = 3, 4, 5$, to compute the hits or initial alignment clues. The k-tuple matching between two sequences is computed in this paper as:

$$T(h,l) = \sum_{j=1}^{k}\sum_{i=1}^{n} \mid X_i(u_{h+j-1}) - X_i(v_{l+j-1}) \mid \tag{7}$$

2.1 Multidimensional Scaling

A problem which must be solved is how compute $D_X(a,b)$ as a good approximation of $d(a,b)$; this requires the computing of the vector set: $\mathbf{X}(a), a \in \mathcal{A}$. The Sammon method [20] is used to achieve this goal; it provides a good ratio of result quality to computational complexity[16–18]. It maps a distance function to a reduced dimensionality space based on the minimization of an objective function assigning to each amino acid tentative coordinates. These coordinates are meaning less, and they are useful only to compute the distance. The Sammon method is based on the minimization of a non-lineal goal function related with the error between the original distances and the tentative ones, consequently several solutions can be obtained if some local minimum exists. The procedure

requires the minimization of the goal function $S(X)$ which can be assimilated to a relative error of the mapping process:

$$\min_X S(X) = \frac{\sum_a \sum_{b<a} \frac{[D_X(a,b)-d(a,b)]^2}{d(a,b)}}{\sum_a \sum_{b<a} d(a,b)} \tag{8}$$

while the relative error is compute as:

$$E(X) = \frac{2}{N(N-1)} \sum_a \sum_{b<a} \frac{[D_X(a,b) - d(a,b)]^2}{d^2(a,b)} \tag{9}$$

where N is the amino acid number. The $\mathbf{X}$ solution is not unique due to the geometrical transformations that preserve the distance D_X. For the L_1 metric the freedom degrees are less that in euclidean or L_2 metric, because the rotation group is finite dimensional in the first case instead of infinite dimensional of the second case. The vector $\mathbf{X}(a)$ provided by the optimization procedure is transformed to the $\mathbf{Y}(a)$ vector in the byte values range $[0, 255]$ by geometrical transformations of translation and scaling. Table 1 contains the second coordinate type for 1,2,4 and 8-dimensional mapping. Due to the hardware restrictions these dimensional values are the most useful for practical proposes. The translation to the origin of coordinates does not modify the distances, whereas the scaling to fit the $[0, 255]$ range modifies the distance with a constant factor ρ related with the scaling transformation. The relation between the distances computed by mean of the two vector type is:

$$D_X(a,b) = \rho D_Y(a,b) \tag{10}$$

3 Results

Both Genetic and Gradient optimization methods can be used to achieve the minimization of the goal function. Gradient procedures have better convergence around local minima, while Genetic procedures allow a better global optimization by considering several local minima. Many solutions are expected in the proposed problem, covering a wide range of both local minimum due to non-linearity and also due to geometrical transformations.

A Genetic Algorithm is used to obtain a solution which is afterward refined by applying a Gradient procedure based on Quasi-Newton algorithm. Genetic algorithm are good to jump far of tentative local minima. However, in practice after a number of iterations the genetic algorithm is mainly working in the refinement of a local minimum, but for this task the gradient procedures are more efficient. The minimum of several trial cases of genetic and gradient procedures is chosen as the solution. GAOT[21] is a public domain Genetic Toolbox that is used for the first stage and the MATLAB Optimization Toolbox[22] for the second one. Figure 1 shows the graphical representation of the value $S(X)$ of the Sammon function and the relative error $E(X)$ vs the dimensionality n of

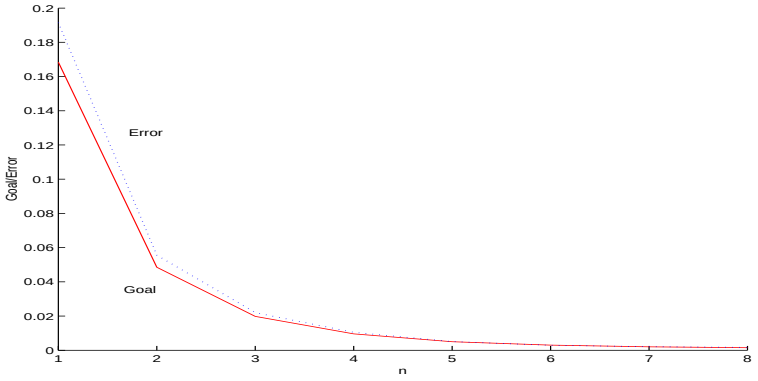

Fig. 1. Goal function $S(X)$ and error $E(X)$ vs the mapping space dimensionality n.

the mapping space. Table 1 contains the obtained $\mathbf{Y}$ coordinates for 1,2,4, and 8-dimensionality.

To illustrate the pattern matching procedure an example with two protein sequences is used. These proteins have the entry names GTH2_TOBAC and GTH1_MAIZE in SWISS-PROT database; both are related proteins, member of the Glutathione S-transferases family[23], included in the GST_C entry of the Pfam protein families database[24].

Figure 2 at left shows the standard dotplot representation of both proteins. The dotplot is the simplest matching procedure, it is a 1-tuple matching. In this figure each point means a score value greater that a threshold. In this case $s(a,b) \geq 4$ according with the BLOSUM65 matrix. The previous alignment of both sequence shows a significative match in the 49-75 region. Other matches are too weak to be considered. Also, Figure 2 at right shows the solution of the matching procedure with tuple size $k = 4$, the mapping dimensionality $n = 2$, by using a threshold $D_Y(U,V) \leq 20$. As shown the significative region is detected as can be supplied to next stages of heuristics procedures as FASTA or BLAST.

Figure 3 shows a comparative evaluation of the computational time of some matching procedures. The sequence of the protein GTH1_MAIZE is matched with some randomly chosen sequences in the SWISS-PROT database. The length of the GTH1_MAIZE sequence is 213 amino acids, the figure shows the computational cost in msec. of each protein match vs the sequence length. To avoid the noise produced by the operating system interruptions and services, no other user task was running and each represented value is the mean over a thousand cases. The computation of the dotplot is compared with the computation of the matching procedure defined in equation (7) with $k = 4$ and $n = 2$; the latter is computed by coding in C language and also by using the MMX instruction set in assembler language. The processor used is a Intel Pentium IV at 2Ghz. It is concluded that the 4-tuple matching coded in MMX has similar cost that the dotplot, but the quality of results is better as shown in Figure 2. The MMX

Table 1. Mapping coordinates for 1,2,4 and 8-dimensionality of BLOSUM62 transformed to integer [0,255] range for use in fast matching procedures

	$n=1$	$n=2$		$n=4$				$n=8$							
Amino Acid	Y_1	Y_1	Y_2	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
A	140	100	180	182	94	68	49	31	66	120	85	139	105	98	40
R	180	161	113	181	16	108	79	24	89	105	38	0	97	77	24
N	203	198	191	244	26	73	71	94	83	35	46	63	89	90	41
D	216	227	168	255	59	51	98	117	62	82	66	64	175	84	72
C	31	0	166	154	191	61	56	31	87	149	114	100	0	178	63
Q	164	184	146	173	10	65	100	34	68	88	35	36	129	43	50
E	190	177	175	205	36	54	110	13	66	87	53	49	176	85	46
G	228	148	233	230	105	101	61	88	67	88	70	175	63	71	11
H	239	207	92	193	0	0	66	25	67	0	69	40	81	73	126
I	97	63	142	157	65	50	8	36	34	158	17	115	103	115	80
L	89	76	124	154	50	70	0	22	24	144	4	101	88	106	66
K	173	170	134	181	42	88	123	24	121	109	42	63	132	73	25
M	115	95	127	151	33	81	23	28	41	149	0	95	98	50	56
F	59	91	76	97	52	37	45	0	20	207	61	88	78	75	96
P	255	129	255	167	65	164	97	37	34	56	167	107	150	69	25
S	154	139	171	199	76	74	76	69	77	102	81	88	107	86	47
T	131	113	198	142	85	73	89	54	30	101	61	77	107	150	34
W	0	121	0	0	46	82	61	33	0	255	69	52	44	0	0
Y	71	129	62	117	44	8	63	37	10	184	74	40	79	71	122
V	105	70	153	163	70	55	16	37	34	157	35	121	116	115	64
ρ	0.1444	0.1036		0.0802				0.0421							

procedure is slight faster than the dotplot in long sequence and also slight slower in short sequences. In all case the 4-tuple matching in C is the slower option.

References

1. Attwood, T., Parry-Smith, D.: Introduction to Bioinformatics. Prentice-Hall (1999)
2. Hunter, L.: Artificial Intelligence and Mollecular Biology. MIT Press (1993)
3. Baldi, P., Brunak, S.: Bioinformatics, The Machine Learning Approach. MIT Press (2001)
4. Bik, A., Girkar, M., Grey, P., Tian, X.: Efficient exploitation of parallelism on pentium III and pentium 4 processor-based systems. Intel Technology Journal Q1 (2001) 1–9
5. Hughey, R.: Parallel hardware for sequence comparation and alignment. CABIOS **12** (1996) 473–479
6. Yap, T., Frieder, O., Martino, R.: Parallel computation in biological sequence analysis. IEEE Trans. on Parall. and Distr. Syst. **9** (1998) 1–12
7. Rognes, T., Seeberg, E.: Six-fold speed-up of smith-waterman sequence database searches using parallel processing on common microprocessors. Bioinformatics **16** (2000) 699–706
8. Rognes, T.: Paralign: a parallel sequence algorithm for rapid and sensitive databases searches. Nucleic Acids Research **29** (2001) 1647–1652

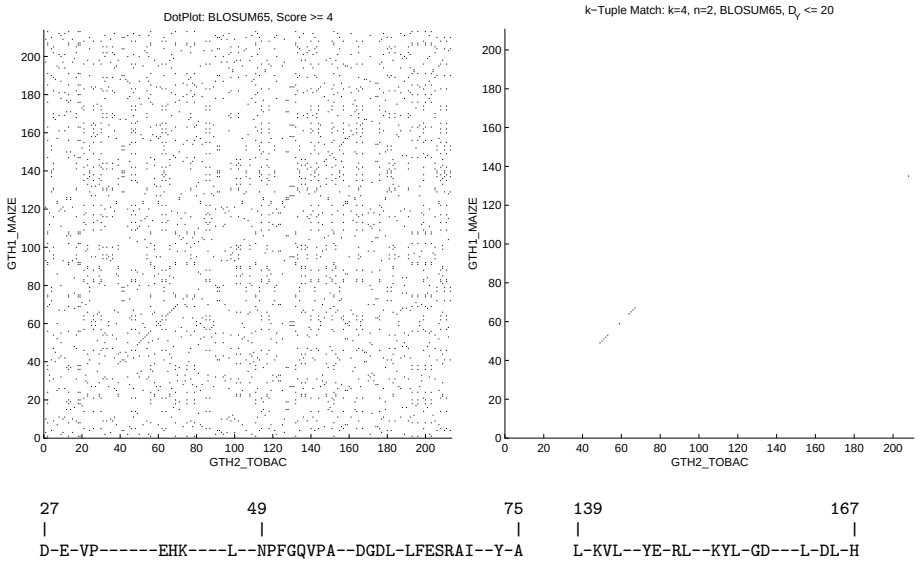

Fig. 2. At left, Dotplot representation between two proteins, GTH2_TOBAC and GTH1_MAIZE. Both are related proteins members of the Glutathione S-transferases family. Each point has a score $s(a, b) \geq 4$ using the BLOSUM62 substitution matrix. The two local alignments between the proteins are shown below with the position in the sequence. Each amino acid symbol means exact match, while the - symbol means mismatch. The stronger similarity is in the 49-75 region, also very week alignments are detected in the 27-48 and 139-167 regions. At right, k-Tuple matching representation between both proteins by using a tuple size $k = 4$ and a mapping dimensionality $n = 2$ of the BLOSUM65 matrix. Shown points have a tuple distance $D_Y(U, V) \leq 10$.

9. Dayhoff, M., Schwartz, R., Orcutt, B.: Atlas of Protein Sequence and Structure. Volume 5. Nat. Biomed. Res. Found. (1978)
10. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. **89** (1992) 10915–10919
11. Smith, T., Waterman, M.: Identification of common molecular subsequences. Jor. Mol. Biol. **147** (1981) 195–197
12. Pearson, W., Lipman, D.: Improved tools for biological sequence comparation. Proc. Natl. Acad. Sci. **85** (1988) 2444–2448
13. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic local aligment search tool. Jor. Mol. Biol. **215** (1990) 403–410
14. de Vel, O., Li, S., Coomans, D.: Non-Linear Dimensionality Reduction: A Comparative Performance Analysis. In: Learning from Data: AI and Statistics. Springer-Verlag (1996) 323–331
15. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley and Sons (2001)
16. Li, S., de Vel, O., Coomans, D.: Comparative performance analysis of non-linear dimensionality reduction methods. Technical report, James Cook Univ. (1995)

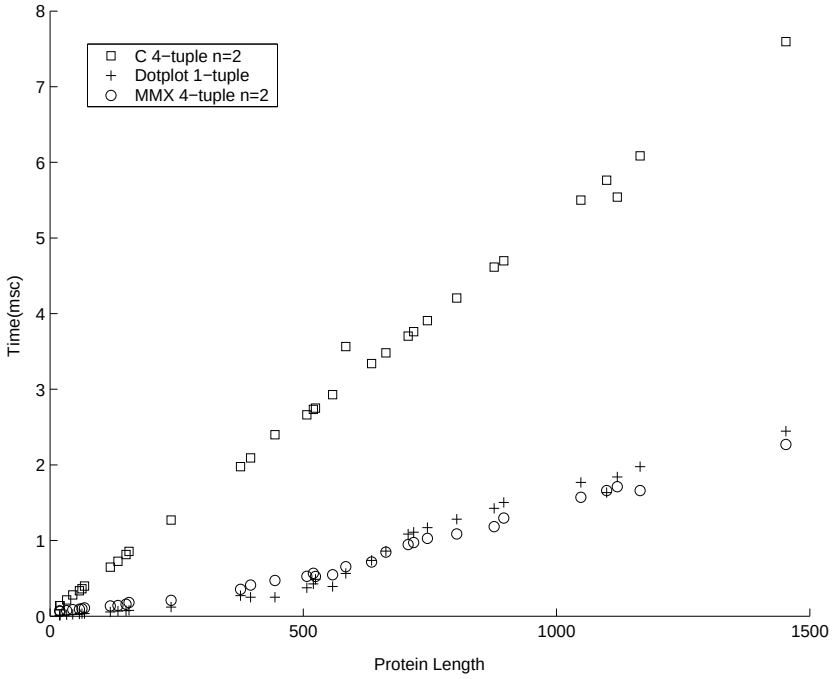

Fig. 3. Computational time in msec. of matching procedures between the GTH1_MAIZE and some randomly chosen protein sequences vs the protein length. Included Procedures are Dotplot, which is an 1-tuple, C and MMX implementations of matching 4-tuple with a mapping dimensionality $n = 2$. MMX implementation has a similar computational cost that Dotplot that is the simplest k-tuple procedure, while it allows a high quality detection of preliminary regions of local alignments.

17. Backer, S.D., Naud, A., Scheunders, P.: Nonlinear dimensionality reduction techniques for unsupervised feature extraction. Pattern Recognition Letters **19** (1998) 711–720
18. Scheunders, P., Backer, S.D., Naud, A.: Non-linear mapping for feature extraction. Lecture notes in computer science **1451** (1998) 823–830
19. Needleman, S., Wunsch, C.: A general method applicable to the search for similarities in amino acid sequences of two proteins. Jor. Mol. Biol. **48** (1970) 443–453
20. Sammon, J.: A nonlinear mapping for data structure analysis. IEEE Trans. Computers **18** (1969) 401–409
21. Houck, C., Joines, J., Kay, M.: A genetic algoritm for function optimization: A matlab implementation. Technical report, NCSU (1995)
22. Coleman, T., Branch, M., Gracce, A.: Optimization Toolbox User's Guide. Mathworks Inc. (1999)
23. Pearson, W.R.: Protein sequence comparation and protein evolution. Technical report, Dept. Biochemistry and Molecular Genetics. Unv. Virginia (2001)
24. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L.: The pfam protein families database. Nucleic Acids Research **28** (2000) 263–266