N-dimensional Mapping of Amino Acid Substitution Matrices

Article · January 2002

CITATION
1

READS
162

3 authors:

Juan Méndez
Universidad de Las Palmas de Gran Canaria
43 PUBLICATIONS **133** CITATIONS
SEE PROFILE

Antonio Falcon
Universidad de Las Palmas de Gran Canaria
30 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Javier Lorenzo-Navarro
Universidad de Las Palmas de Gran Canaria
129 PUBLICATIONS **1,075** CITATIONS
SEE PROFILE

N-Dimensional Mapping of Amino Acid Substitution Matrices

Juan Méndez, Antonio Falcón, and Javier Lorenzo

Instituto de Sistemas Inteligentes. IUSIANI
Univ. Las Palmas de Gran Canaria, Spain
[jmendez,afalcon,jlorenzo]@dis.ulpgc.es

Abstract. A procedure to map score matrices in n-dimensional spaces is presented. Score or substitutions matrices are used as similarity-like measure between amino acid in protein alignment procedures. The first stage of heuristic local alignments procedures as FASTA and BLAST uses local matching of very short sequences, also named k-tuples. By using L_1 metric this matching task can be computed very fast. This procedure can be implemented by using SIMD instructions which are present in most of low cost microprocessors included in personal workstations and server. To design this procedure a table that maps the scores matrices as PAM y BLOSUM are 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 MDS as used in Pattern Recognition and Machine Learning. Previously, a distance function must be defined from the score matrix. To map the distance function a variation of the Sammon non-lineal dimensionality reduction procedure is used with a genetic algorithm that minimizes a goal function. To fit the SIMD constraints, both the dimension k of tuples and the space dimensionality n must verify: $k \times n = 8 \times m$. The table results for the BLOSUM62 with 1,2 and 4-dimensionality and graphical representations of the solution map are included. These last show that the biochemical amino acid groups are well mapped as data cluster; also the strong hydrophobic residues have a highlight spatial property, because there are linearly separable in 2-dimensional mapping.

1 Introduction

The fast growing of information contained in the biological databases[1] requires more efficient processing systems to found 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 techniques as used in Knowledge and Data Engineering, Pattern Recognition and Machine Learning subareas can provide additional approaches to allows better computational performances in Gemomic related systems[2]. This paper uses Pattern Recognition and Machine Learning techniques applied in Bioinformatics[3] to obtain data tables needed 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 the microprocessors of most the low cost computer systems, as Intel Pentium 4 and AMD K4. 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 these are design according the above processor characteristics[4].

Some works are dealing with the use of parallel computation for sequence analysis[5, 6], and also in the use of SIMD instructions in the improvements of local alignments[7, 8]. However, this work does not deal with hardware or architectural proposal, it 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 that 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 simplistic ASCII code matching, a n-dimensional code matching is proposed based on the biological information contained in the score o substitution matrices. The information retrieval procedure take advantage of two architectural improvements of modern microprocessors: parallel computation with multiple data processing units, and sequential memory access which increases the cache throughput. Rather than dealing with computer architectural issues or SIMD programming, 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 used mainly in Pattern Recognition and Machine Learning for feature selection and also for visualization of high dimensional data sets.

2 Low Dimensional Mapping of Score Matrices

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 Integer SIMD or MMX instruction set included to improve the performance of multimedia, text retrieval and signal processing applications. Genomic related computations can exploit this improvement, but it requires different approaches for some of the actual algorithms. Most of problems related with sequence analysis are based on score matrices to model the amino acid distances and similarities. Perhaps, this is not the best 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 and affinity 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 $s(a, a) > s(a, b)$, also is 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. The first are obtained from observed mutations in general alignments of sequences, and the second from specific substitutions in blocks of aligned sequences contained in the BLOCKS database[19]. The score matrix in this cases is defined as:

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

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 proposed distance is expressed as a probability ratio:

$$d(a,b) = -\frac{2}{\lambda} \log \frac{p(a,b)}{\sqrt{p(a,a)p(b,b)}} \tag{5}$$

The score of a k-tuple of two sequence U and V is computed in the alignment procedures[11, 20] by using substitution matrices as:

$$s(U,V) = \sum_{j=1}^{k} s(u_j, v_j) \tag{6}$$

Where $u(j)$ and $v(j)$ are the amino acid in the k-tuple. If the distance of this k-tuple, $d(U,V)$, is defined as:

$$d(U,V) = s(U,U) + s(V,V) - 2s(U,V) \tag{7}$$

It can be computed as:

$$d(U,V) = \sum_{j=1}^{k} d(u_j, v_j) \tag{8}$$

If $d(a,b)$ can be computed by $D_X(a,b)$ with a reduced error, then the computing of $d(U,V)$ can be achieved by the expression:

$$D_X(U,V) = \sum_{j=1}^{k} \sum_{i=1}^{n} \mid X_i(u_j) - X_i(v_j) \mid \tag{9}$$

which 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 is lower significative; BLAST uses $k = 3, 4, 5$, to compute the hits or initial alignment clues.

The 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 [21] 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 by assign 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 exist. A slight modification of the Sammon method is used to adapt it to the biological environment. The procedure requires the minimization of the goal function $E(X)$ which can be assimilated to a relative error of the mapping process:

$$\min_X E(X) = \frac{\sum_a \sum_{b<a} p_{ab} \frac{(D_X(a,b) - d(a,b))^2}{d(a,b)}}{\sum_a \sum_{b<a} p_{ab} d(a,b)} \tag{10}$$

Where is introduced $p_{ab} = p(a)p(b)$ as the probability of the residue pair, included to reflect the different frequency of residues in protein sequences. The $\mathbf{X}$ solution is not unique due to the geometrical

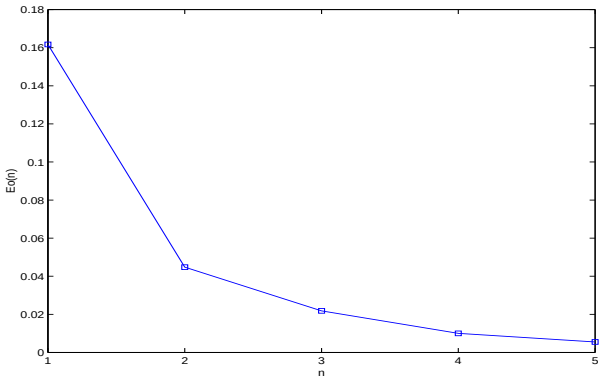

Fig. 1. The minimum value E_0 of the Sammon function for BLOSUM62 vs the dimensionality n of the mapping space. A fast convergence is found for low dimensionality, because the E_0 value can be considered as the relative error of the mapping process.

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.

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 across far 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[22] a public domain Genetic Toolbox is used for the first stage and the MATLAB Optimization Toolbox[23] to the second one. Results for the mapping process with $n = 1, \ldots, 5$ have been obtained for the BLOSUM62 matrix. The Figure 1 shows the graphical representation of the value E_0 of the Sammon function $E(\mathbf{X})$ in the best obtained minimum, which are: $(0.1617, 0.0448, 0.0218, 0.0100, 0.0055)$. These vales, which can be considered as relative error of the mapping process, verify a fast convergence to near null error at relatively low dimensionality.

The amino acid coordinates for 1,2,4-dimensional mapping are included in the Table 1. Due to the hardware restrictions these dimensional values are the most useful for practical proposes. The 1-dimensional can be used for low precision but high speed matching, while the 4-dimensional with 4-tuples matching can be considered a optimal solution with a good precision and near the the tuple size of FASTA and BLAST. The 2-dimensional is a intermediate case which also allows practical visualizations as shown in the Figure 2.

The substitution properties contained in the score matrix are represented by means of a symmetrical set of numerical values; when this information is mapped in a plane space the biological affinity between the amino acid is more clearly shown. Some biochemical groups are nearly mapped conforming significative clusters as the aromatic(F,Y,W,H), the basic or positive charged(H,K,R) and the aliphatic(V,I,L).

A data transformation is needed to use the information contained in the coordinate table for practical matching procedures that use the hardware speed up capabilities of popular microprocessor. The vector $\mathbf{X}(a)$ provided by the optimization procedure are transformed to the $\mathbf{Y}(a)$ vector in the byte values range $[0, 255]$ by geometrical transformations of translation and scaling. The Table 2 contains the second

Table 1. Mapping coordinates for 1,2 and 4-dimensionality reduction of BLOSUM62 substitution matrix. P is the prior probability of amino acid, and E_0 is the value of the Sammon function in the local minimum.

			$n=1$	$n=2$		$n=4$			
		P	X_1	X_1	X_2	X_1	X_2	X_3	X_4
Ala	A	0.10	19.646	25.829	20.349	18.494	21.027	14.399	18.471
Arg	R	0.05	15.773	17.835	25.406	19.704	18.823	14.478	26.680
Asn	N	0.04	9.986	29.136	26.147	22.539	19.199	11.843	22.854
Asp	D	0.06	8.167	26.679	30.911	12.585	18.106	11.964	21.154
Cys	C	0.01	35.376	34.618	17.522	25.835	14.300	16.941	19.760
Gln	Q	0.04	13.367	21.224	27.177	15.245	18.846	15.614	22.733
Glu	E	0.06	11.824	23.695	29.288	15.671	17.912	13.505	23.308
Gly	G	0.08	6.073	32.288	23.165	19.064	24.233	10.820	21.189
His	H	0.02	3.934	12.508	26.142	14.642	20.508	17.885	25.931
Ile	I	0.06	26.608	22.769	14.798	19.601	18.124	18.299	16.729
Leu	L	0.09	28.163	20.497	16.405	20.629	18.094	19.044	18.725
Lys	K	0.06	14.527	19.767	27.697	18.375	17.090	14.776	25.064
Met	M	0.02	24.139	19.266	18.065	21.854	18.412	17.694	20.511
Phe	F	0.04	31.968	14.880	18.646	19.652	20.213	22.881	20.750
Pro	P	0.04	3.032	22.945	34.379	17.628	10.524	14.888	20.956
Ser	S	0.06	17.249	25.113	24.127	17.864	19.327	12.540	19.963
Thr	T	0.06	21.653	22.308	22.132	19.628	15.473	12.956	19.083
Trp	W	0.01	39.000	6.482	22.421	14.012	22.970	17.177	12.287
Tyr	Y	0.03	32.946	12.977	21.033	17.782	19.568	23.794	21.770
Val	V	0.07	25.157	23.557	15.614	19.047	17.879	17.248	16.878
	E_0		0.1617	0.0448		0.0100			

coordinate type. The translation to the origin of coordinates do not modifies the distances, whereas the scaling to fit the $[0, 255]$ range modifies the distance with a constant factor μ related withe the scaling transformation. The relation between the distances computed by mean of the two vector type are related as:

$$D_Y(a, b) = \mu D_X(a, b) \tag{11}$$

The Figure 3 shows the representation in the new coordinates system $\mathbf{Y}$. In this Figure, and also in the previous one, can be show as the strong hydrophobic group(WYFVILM) is mapped in the lower left corner of the map, linearly separated from the other amino acid. This fact provides qualitative evidence about the existence of some relation between the mapping space and the hydrophobic property. This connects with a more general problem related with the possible biological and biochemical meaning of the axes obtained in the mapping processes.

The hydrophobic/hydrophilic properties of amino acid are fundamental items in the dynamics and structure of proteins[24]. Due that the biological matter is basically an aqueous solution, the water affinity is essential in the relation of a protein with its environment. The mutations with significative changes in the water affinity have a high probability of generate disfunctions, so having a low probability of survivance, therefore being lost in the evolution process. Survival mutations in stable species are the observed and coded in biological databases and translated to the score matrices.

There are many hydrophobicity, or its inverse hydrophilicity, scales for amino acid residues in proteins[25–28]. Unfortunately most are uncorrelated and they self contradictory. One of most uses is the Levitt hydrophobicity scale[26] which has been used to make a score matrix for protein alignment[29]. The Figure 3 shows the relation between the Levitt scale and the coordinate in a 1-dimensional map. In this case exists a high correlation between the meaning-less X coordinate and the scale for the strong hydrophobic group. This agrees to the hypothesis that some semantical and meaning can be discovery in the axes of the mapping process; as initial clue, for BLOSUM62 exist some relation of the mapping axes with the amino acid hydrophobicity. However the main problem with this hypothesis comes with the ambiguous numerical definition of hydrophobicity due to the proliferation of uncorrelated scales.

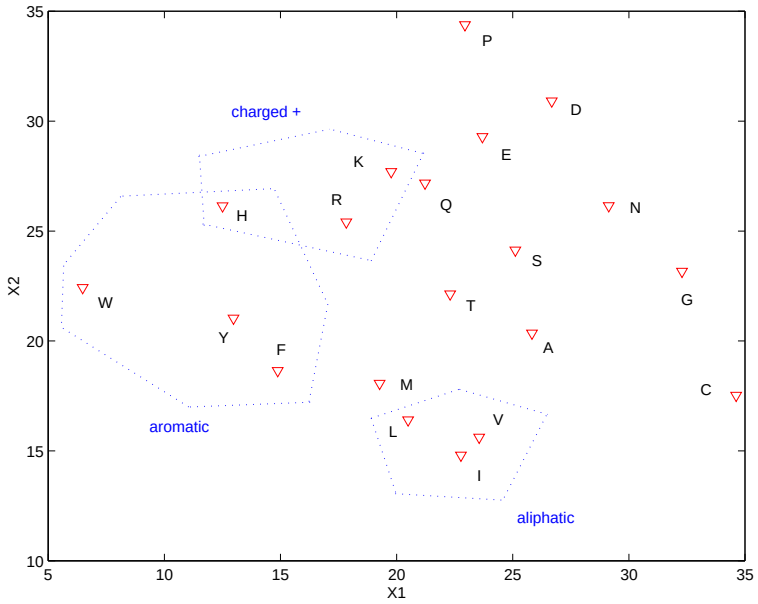

Fig. 2. 2-Dimensional Mapping of Amino Acids based on BLOSUM62 matrix. Some biochemical groups are outstands to show that are near mapped according with their biological affinity

From the obtained results, further works are required in two far areas. The first in the implementation of fast matching procedures that uses the proposed tables; this is mainly concerning with software efficiency in the use of microprocessor improvements. The second is a more wide and intensive study of the relation between the mapping coordinates and the hydrophobicity and other semantic concepts.

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 axploitation 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
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. Jhon Wiley and Sons (2001)

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

	$n=1$	$n=2$		$n=4$			
	Y_1	Y_1	Y_2	Y_1	Y_2	Y_3	Y_4
A	118	175	50	105	186	63	110
R	90	103	96	126	147	65	255
N	49	205	103	176	154	18	187
D	36	183	146	0	134	20	157
C	229	255	25	235	67	108	132
Q	73	134	112	47	147	85	185
E	62	156	131	55	131	48	195
G	22	234	76	115	243	0	158
H	6	55	103	36	177	125	242
I	167	148	0	124	135	133	79
L	178	127	15	143	134	146	114
K	81	120	117	103	116	70	226
M	150	116	30	164	140	122	146
F	205	76	35	125	172	214	150
P	0	149	177	89	0	72	154
S	101	169	85	94	156	30	136
T	132	143	66	125	88	38	120
W	255	0	69	25	221	113	0
Y	212	59	57	92	160	230	168
V	157	155	7	114	130	114	81
μ	7.0896	9.0632		17.7174			

16. Li, S., de Vel, O., Coomans, D.: Comparative performance analysis of non-linear dimensionality reduction methods. Technical report, James Cook Univ. (1995)
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. Henikoff, S., Pietrokovski, S., Henikoff, J.: Superior performance in protein homology detection with the blocks database servers. Nucleic Acids Research **26** (1998) 309–312
20. 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
21. Sammon, J.: A nonlinear mapping for data structure analysis. IEEE Trans. Computers **18** (1969) 401–409
22. Houck, C., Joines, J., Kay, M.: A genetic algoritm for function optimization: A matlab implementation. Technical report, NCSU (1995)
23. Coleman, T., Branch, M., Gracce, A.: Optimization Toolbox User's Guide. Mathworks Inc. (1999)
24. Gerstein, M., Levitt, M.: Simulating water and the molecules of life. Scientific American (1998) 100–105
25. Cornette, J., Cease, K., Margalit, H., Spouge, J., Berzofsy, J., DeLisi, C.: Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J. Mol. Biol. **195** (1987) 659–685
26. Levitt, M.: A simplified representation of protein conformations for rapid simulation of protein folfing. J. Mol. Biol. **104** (1976) 59–107
27. Kyte, J., Doolittle, R.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157** (1982) 105132
28. Karplus, P.: Hydrophobicity regained. Protein Sci **6** (1997) 1302–1307
29. George, D., Barker, W., Hunt, L.: Mutation data matrix and its uses. Methods Enzymol. **183** (1990) 333–351

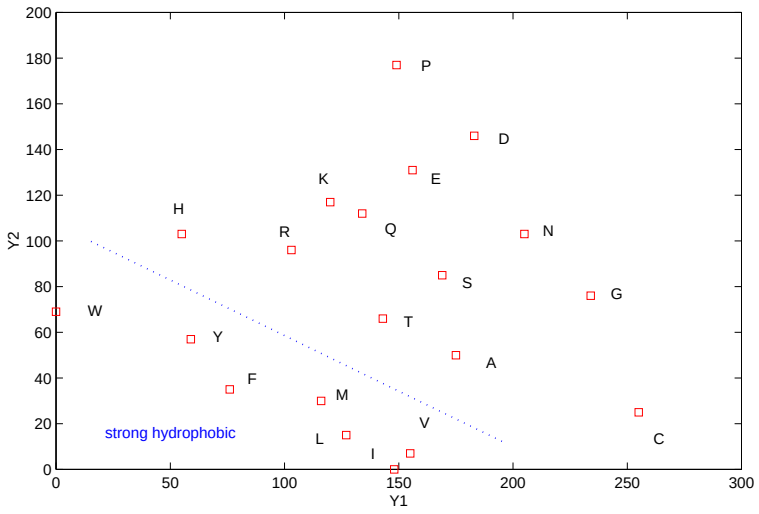

Fig. 3. 2-Dimensional Mapping of BLOSUM62 matrix with Y coordinates showing the strong hydrophobic group which can be linearly separable.

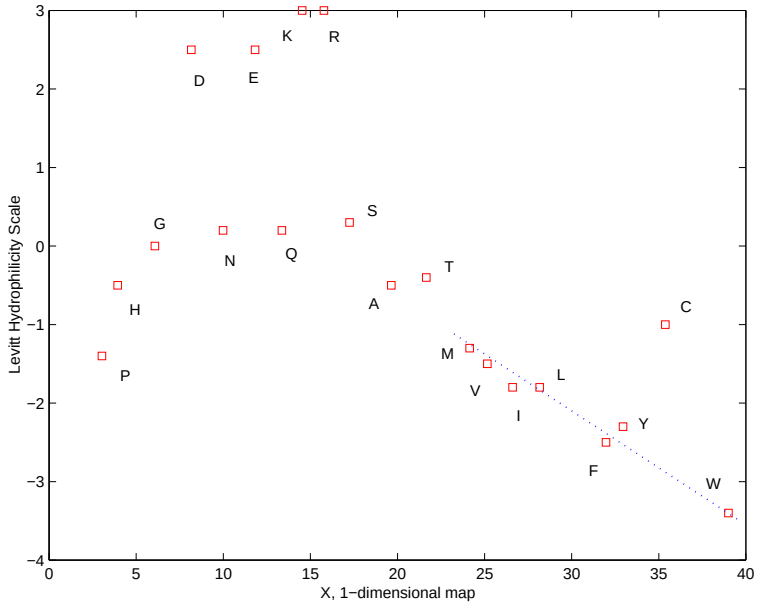

Fig. 4. The 1-dimensional map vs the Levitt hydrophilicity scale showing a high correlation degree for the strong hydrophobic residues.