ELSEVIER

Peptides quantitative structure–function relationships: An automated mutation strategy to design peptides and pseudopeptides from substitution matrices

M. Adenot, C. Sarrauste de Menthière, A. Chavanieu, B. Calas, and G. Grassy

Centre de Biochimie Structurale, CNRS UMR 9955, INSERM U 414, Faculté de Pharmacie 15, Avenue Charles Flahault, 34 060 Montpellier Cedex, France

The process by which analogs in peptide chemistry are currently designed does not include any quantitative basis for amino acid substitutions from pharmacological leads. Here, we show that substitution matrices such as PAM 250 can provide quantitative constraints compatible with biological activity. This article describes its use in a strategy of rational amino acid substitution in peptides and proteins: we have computed a chemically derived matrix equivalent to the well-known PAM 250 matrix, reflecting the natural mutability rates of amino acids in protein evolutions but that can be extended to all the noncoded amino acids. Some of these noncoded amino acids are widely used to mimic secondary structure, to constrain backbone conformation, or to evade protease degradation. An automated sequence mutation (ASM) strategy has been defined to generate mutations within constraints. Application of such a substitution matrix to quantitative structure–function relationship studies will be of use in the design of proteins and peptides destined to become pharmaceutical drugs. In particular, issues such as which functionally conserved substitutions are able to satisfy conformational restrictions, oral bioavailability, or formulation demands can be quantitatively addressed. © 2000 by Elsevier Science Inc.

Corresponding author: M. Adenot, Centre de Biochimie Structurale, 40 Rue Antoine Lumiere, 69008 Lyon, France.

Received 24 May 1999; revised 15 November 1999; accepted 18 November 1999.

Keywords: substitution matrices, distance mapping, quantitative structure–function relationships, protein engineering

INTRODUCTION

Peptide sequence modulation searches for the substitutions that maximize similarities within experimental constraints. Many attempts to describe similarity between amino acids have been made since Sneath,[1] with many applications to problems such as sequence alignments. It has been shown that the chemical differences between residues can be linearly related to the substitution frequencies in related proteins.[2] However, it seems that there is no correlation between substitution frequency and physicochemical difference in abnormal proteins or in certain regions of proteins characterized by rapid rates of substitution. This reflects a process of mutation, subject to random exploration of genetic diversity, rather than natural selection, which always leads to protein functional conservation or optimization. Drug design can also be considered to be a combination of innovation and optimization steps, leading to a formal analogy with molecular evolution. The random exploration of molecular diversity is analogous to the lead generation step in combinatorial chemistry. On the other hand, the natural selection process is rather similar to the drug optimization step: it searches for conservation of biological function with an adaptation of the global properties to a given environment, introducing adaptive constraints. In fact, with peptide design, chemists often search for potent peptide analogs that are metabolically stable or that fit some structural or physicochemical constraints. Genetic algorithms have been designed that generate molecular structures within constraints,[3] based on the principles of Darwinian evolution. The main goal of evolutionary computation is the *in*

Journal of Molecular Graphics and Modelling 17, 292–309, 1999

655 Avenue of the Americas, New York, NY 10010

1093-3263/99/$–see front matter
PIIS1093-3263(99)00037-6

silico application of the concepts of natural selection to a population of structures.[4]

Substitution matrices, widely used for sequence alignements, give a basis for simulation of the molecular evolution process and could lead to peptide analogs phylogenetically related to a peptide leader. These matrices are designed for alignments of sequences containing the 20 natural amino acids and rely on observation of mutation rates in families of related proteins[5] or blocks of aligned sequence segments.[6]

Since peptide chemistry make use of hundreds of noncoded amino acids, it was necessary to design substitution matrices on the basis of various physicochemical, topological, or structural properties. Many authors have proposed chemically or structurally derived matrices in order to detect the properties that explain the evolutionary driving force or to measure amino acid dissimilarities. The traditional qualitative approach has been pioneered by Sneath,[1] who assembled a dissimilarity matrix consisting of 134 qualitative variables related to the presence or absence of functional groups and gap values of various physicochemical properties such as p*K* at isoelectric point, optical rotation in HCl, or R_f in paper chromatography. Quantitative descriptions of amino acid similarity have been proposed by different authors with physicochemical descriptors such as polarity, volume, size, composition,[7,8] hydrophobicity,[9] topological[10,11] or structural descriptors.[12–14] Each of these matrices can be used for a particular task, according to the kind of properties the user wishes to conserve in the substitution. It is generally accepted that amino acid residues at functional sites are highly conserved in protein evolution.[15] Many authors concede that the fundamental properties necessary for conservation of activity or protein folding are related to lipophilicity and volume.[2,7,16]

Difficulties often arise from the fact that chemists want both to conserve the overall shape of essential residues and eliminate certain undesirable features such as susceptibility to proteases and low bioavailability. Sometimes, it is necessary to introduce conformational constraints using β-turn or γ-turn mimics, constrained amino acids, or functional residues for side-chain cyclization or attachment of drugs. Peptide chemistry today makes use of a wide range of amino acid derivatives that exhibit such characteristics, but choosing the best amino acid derivative is by far the most difficult aspect of peptide pharmacology. In this study, we attempt to give a quantitative rational basis for the substitution of replaceable residues in the design process of peptide drug candidates. Since distance matrices define the similarity between amino acid derivatives in terms of chosen physicochemical properties, we propose to use them as a tool for peptide quantitative structure–function relationships and generation of potentially biologically active peptides. The basic assumption is that, for a given chemical series, the activity is a continuous variable in descriptor space and thus the optimal compound is located necessarily in the vicinity of the leader compound. Exploration of this parameter space should define the limits of the activity range and, by definition, identify the ideal molecules.

METHODOLOGY

Derivation of a Substitution Matrix Aligned on PAM 250 Substitution Matrix

Molecular description of the chemical diversity of an amino acid fragment bank Whatever the method employed, one critical step of this design process is the choice of a set of noncorrelated (orthogonal) descriptors that can adequately describe the physicochemical space of the substituent fragment bank. Sneath[1] and other authors have attempted to describe qualitatively or quantitatively the structural and physicochemical shape of the 20 natural amino acids. Jonsson et al.[17] extended this work for 35 nonnatural amino acids that are frequently used in peptide chemistry. A complete description of the amino acids was given by experimental determination of their physicochemical properties. This huge amount of descriptor data is generally reduced by calculating only the first principal components of the principal component analysis (PCA) and using them as descriptors of the overall structural and physicochemical similarity of the amino acids. Hellberg et al.[18] found that the first component is mainly related to hydrophilicity, the second is influenced by the size, and the third by electronic properties. Similarly, Bogardt et al.[19] have shown that only 4 of 16 descriptors were sufficient to explain 96% of the total variance of amino acid properties.

There is a major drawback in using nonspecific descriptors as principal components: they give a measure of global chemical similarity but do not emphasize the one or two single properties that explain activity at a molecular level. Such methods cannot identify the critical determinant of each initial property that relates to activity. In contrast, where distances are calculated over one or more properties, the influence of each property individually, or combined, can be tested. This is essential to understand the physical nature of the molecular recognition process and to derive rules for design of new active molecules.

Because of the current use of nonnatural amino acids in the study of peptide structure–activity relationships and in peptide pharmacomodulation, it is necessary that all descriptors should be computable *ex nihilo*. More than 5 000 amino acid compounds are commercially available and some 300 among them occur naturally.[20] In this study, a trial set of 57 noncoded residues have been included in an extended database of amino acids.

We quantitatively described each amino acid by 40 initial descriptors, listed in Table 1, which characterize lipophilicity, molecular geometry, topology, and electronic properties. On the basis of the correlation matrix, only 7 noncorrelated or weakly correlated ($r < 0.65$) descriptors were finally retained.

The lipophilicity (LIP) is described by logP.[21] An additional term for carboxyl- and sulfur-containing amino acids leads to a corrected LIP that fits exactly the experimental polarizability values (POL) of Grantham[7]

$$\mathrm{POL} = \mathrm{LIP} - 2N_S - N_{COO^-}$$

N_S and N_{COO^-} are, respectively, the number of sulfur atoms and carboxyl moieties in the molecule. In many cases, POL will give better results than LIP itself. The geometry is represented by molecular volume (VOL) and the length of the second principal axis of the inertial ellipsoid of the ellipsoidal volume (RYI). RYI is the only principal axis that is not correlated with the molecular volume. The topology is characterized by the Kappa3 connectivity index (KAP),[22] which measures the level of branching at the center of the molecule, and by the Balaban index (BAL),[23] which reflects the molecular compactness. The electronic factors are the lowest unoccupied molecular orbital (LUM), characterizing the electron acceptor

Table 1. Forty Initial Descriptors of the 20 Natural Amino Acids

1. Physicochemical descriptors
1 Lipophilicity
2 Molecular volume
3 Molar refractivity
4 LUMO energy level
5 HOMO energy level
6 Total dipole moment
7 Heat of formation
2. Topological descriptors
8–18 2D VAC components
19 Wiener index
20 Randic index
21 Balaban index
22 Kier ChiV0
23 Kier ChiV1 (Path)
24 Kier ChiV2 (Path)
25 Kier ChiV3 (Cluster)
26 Kier ChiV4 (Path/Cluster)
27 Kappa 1
28 Kappa alpha 1
29 Kappa 2
30 Kappa alpha 2
31 Kappa 3
32 Kappa alpha 3
3. Geometric descriptors
33–35 *X*, *Y*, and *Z* inertia moment components
36–38 *X*, *Y*, and *Z* axis lengths
39 Ellipsoidal volume
40 Surface area

character of the residue, and the dipole moment (DIP), which measures the polar character. Partial charges were previously computed by the semiempirical quantum mechanical AM1 method. This set of descriptors covers structural and topological, as well as physicochemical, diversity of the amino acids.

Calculation of 127 Euclidian distance matrices From the seven previous amino acid descriptors, we derived $\sum_{i=1}^{7} C_i^7 = 127$ distance matrices corresponding to the various combinations of these descriptors with one another. For each matrix, 400 Euclidian distances between the 20 natural amino acids were calculated in an *n* descriptor space, after the initial data were scaled by mean and standard deviation. The Euclidian distance between two amino acids *A* and *B* is calculated as follows:

$$D_{A\rightarrow B} = \sqrt{\sum [X_i(A) - X_i(B)]^2}$$

for the X_1 to X_n standardized descriptors. This technique, known as nearest neighbor analysis, provides a quick and easy method of spanning similarity, according to a desired set of criteria.

Calculation of 14 similarity matrices Molecular similarity indices can be calculated by numerical integration within the ASP program[24] over a 3D grid surrounding both molecules to be compared.

$$C_{AB} = \sum_{i=1}^{N} (\rho_A \rho_B) \Big/ \left(\sum_{i=1}^{N} \rho_A^2 \right)^{1/2} \left(\sum_{i=1}^{N} \rho_B^2 \right)^{1/2}$$

where *N* is the total number of grid points and ρ_A is the matched property (electrostatic potential or lipophilicity) for compound *A* at the *i*th grid point. This Carbo index (C_{AB}) takes a value of +1 when the two compounds have similar properties and a value of −1 when they have totally dissimilar properties.

The Carbo index can be adapted to measure similarity in terms of molecular shape.[25] In this case, the previous equation can be written

$$C_{AB} = N_{AB}/(N_A N_B)^{1/2}$$

where N_{AB} is the number of grid points lying within both superimposed molecules and N_i is the number of grid points lying within molecule *i* only. The optimization uses translations and rotations of the molecules but does not involve conformational changes. Various combined indices can be calculated by linear combination of single property indices. In a first series, the molecules were prealigned using moments of inertia. In a second series, the molecules were prealigned using size, electrostatic potential, and lipophilicity. For both series, each of the 400 pairs of residues was characterized by three Carbo indices derived from single properties (electrostatic potential, lipophilicity and molecular shape respectively) and four combined Carbo indices. These calculations lead to 14 additional Carbo similarity matrices without any bearing on the 127 previous distance matrices. Since Euclidian distances are positive scores and measure dissimilarities, the Carbo similarity indices were transformed:

$$S_{AB} = 1 - C_{AB}$$

where S_{AB} is a positive substitution score, in the range from 0 to 1, comparable with the Euclidian distance. At this point, 140 substitution matrices can be interrogated.

Correlation between PAM 250 and the 140 substitution matrices One of the most widely used similarity measures in protein sequences is the substitution matrix PAM 250 calculated by Dayhoff et al.[5] and extensively used in heuristic algorithms of sequence alignment. It was calculated on the basis of nearly 1 600 accepted point mutations in 71 groups of closely related proteins (<15% different). One of the disadvantages of such matrices in protein alignments is their inability to detect similarity between distantly related sequences. Improved models derived from either sequence-based alignments, such as BLOSUM 62, or structure-based alignments, such as the structural superposition matrix of Risler et al.,[14] perform much better in detecting similarity between distantly related proteins.

For peptide design purposes, the situation is quite different because the chemist or pharmacologist is dealing with closely related sequences in a given test series. The mutation strategy is an attempt to explore the similarity space surrounding the leader compound and to discover drug candidates that conserve the biological function. For this situation, the PAM evolutionary model of Dayhoff et al. will be effective since it is required to simulate mutations at a relatively close evolutionary. It is

well known that the performance of matrices based on the PAM evolutionary model decreases with increasing evolutionary distances. Indeed, at large evolutionary distances (>400 PAM), the matrix values exhibit an asymptotic behavior and transitions to any residue become equal to their frequency of occurrence.[26] The scores do not discriminate sufficiently for further quantitative structure–function relationship analysis. Furthermore, at close distance, the diagonal term of the matrix is largely predominant, leading to a low variability in mutant sequences. For instance, at 100 PAM, the only significant mutation for most amino acid consists of the replacement of the target residue by itself, corresponding to a test series of null variability. These considerations imply that mutations must be simulated for an optimal mean evolutionary distance: the PAM 250 matrix has been retained here as a reference matrix to describe the evolutionary similarity of the side chains of amino acids.

A substitution matrix column constitutes a mutability shape that characterizes each amino acid residue. We calculated column-to-column correlations between the different selected distance matrices and the widely used PAM 250 matrix. For each residue, we selected the substitution shape that is better correlated to the corresponding PAM 250 mutability shape (Table 2). The resulting hybrid distance matrix is called DISMAT 250 (Table 3). It consists in a chemically derived matrix equivalent to PAM 250 in which all the amino acid distances are not evaluated using a common descriptor space. As a consequence, the distances do not appear to be invariant to order. The matrix can be extended to noncoded amino acids and provide a rational basis for choice of substitution at every position of a peptide or artificial protein. In particular, all noncoded amino acid structures that are widely used in pharmacomodulation can be chosen using the same criteria that nature might use if they were coded amino acids. It is not surprising that properties such as lipophilicity and molecular volume are highly concerned, as pointed out by several authors.[16,27,28] The molecular topology was found to be significant for 7 residue types among 20 (Ile, Leu, Met, Trp, Tyr, and Val). The percentage of PAM 250 variance explained by DISMAT 250 is high for most residues, except Arg, Cys, and Met, which are poorly represented in this model. Thirteen residues exhibit more than 75% of the variance explained by the model.

Alignment performances of DISMAT 250 Since the mutability shape of PAM 250 is properly fitted by DISMAT 250 scores, we should obtain similar performances in alignment of protein sequences. A linear inversion of DISMAT 250 scores (S_{ij}) is necessary to transform the distance dissimilarity matrix into a similarity matrix.

$$S'_{ij} = 1 - (S_{ij}/\max(j))$$

The S'_{ij} scores of the intermediate matrix are in the [0; 1] range. Scaling permits fitting the matrix to PAM 250 diagonal score values. The substitution matrix is called FITMAT 250, indicating that DISMAT 250 has been fitted to PAM 250 (Table 4),

$$S_{ij}(\text{FITMAT 250}) = \text{integer}[A(j)S'_{ij} - \max(j)]$$

where $A = \max(j) - \min(j)$ is the jth PAM 250 columns amplitude. FITMAT 250, shown in Table 4, is analogous to a log-odd matrix, even if it is not symmetrical, while substitution scores in a column can be interpreted in terms of the probability that a given $i \rightarrow j$ substitution is more or less frequent than in random sequences of the same composition. As pointed out by Altschul,[29] any matrix of values used for scoring alignments is a log-odd matrix, whatever the model underlying the scores calculation. In fact, FITMAT 250 is mathematically comparable to any matrix based on the PAM evolutionary model. Identities and conservative replacements have positive scores, while unlikely replacements have negative scores. The distribution of scores into a column constitutes a specific mutability shape, which characterizes each of the 20 natural amino acids.

Since the diagonal elements of the FITMAT 250 matrix are larger than the others, the matrix is expected to give satisfactory results in aligning two closely related protein sequences. In a more comprehensive evaluation of the FITMAT 250 matrix performance, we also investigated the ability of FITMAT 250 to detect the similarity between protein sequences with respect to PAM 250.

A strategy of comparison of matrix performance, proposed by Henikoff and Henikoff,[30] has been applied here. BLASTP version 1.4 was used with default parameters ($W = 3$, $E = 10$) to search 257 queries against the SWISS-PROT databank, using FITMAT 250. BLAST is a heuristic search algorithm tailored for sequence similarity searching.[31] Comparison of BLAST results with a list of sequences related to each query permits counting the number of true positives detected by BLAST with both FITMAT 250 and PAM 250. As a control, we used the identity matrix that scores +PAM 250(i, i) for

Table 2. PAM 250 Highly Correlated Substitution Matrices

Amino acid	Substitution matrix	Percentage of explained variance
A	VOL-POL-RYI	82
R	VOL-LIP	58
N	POL	84
D	VOL-POL	88
C	E2[a]	54
Q	VOL-POL	77
E	VOL-POL	79
G	VOL-POL	86
H	VOL-POL	76
I	VOL-LIP-KAP-BAL	75
L	VOL-LIP-KAP	62
K	LS2[b]	60
M	VOL-POL-KAP-BAL	56
F	LSE1[c]	75
P	LSE1[c]	78
S	LS2[b]	74
T	VOL-POL	78
W	VOL-RY-BAL	77
Y	BAL-RY-KAP	77
V	VOL-LIP-BAL	61

[a] Carbo index matrix relative to electrostatic potential *E* (second series).
[b] Carbo index matrix relative to lipophilicity *L* and molecular shape *S* (second series).
[c] Carbo index matrix relative to lipophilicity *L*, molecular shape *S*, and electrostatic potential *E* (first series).

Table 3. DISMAT 250 20 × 20

	A	R	N	D	C	Q	E	G	H	I
A	0	3.042	1.484	1.843	0.992	1.987	2.077	0.714	2.091	2.663
R	2.999	0	0.267	1.975	0.858	1.054	1.364	3.452	1.128	4.105
N	1.743	1.784	0	0.485	0.643	0.736	0.573	1.790	0.679	2.991
D	2.024	1.859	0.647	0	0.889	1.041	0.612	1.951	0.860	2.734
C	1.643	2.210	3.285	2.901	0	2.425	2.806	2.234	2.661	1.952
Q	2.203	1.054	0.252	1.041	0.716	0	0.525	2.416	0.271	3.041
E	2.085	1.312	0.395	0.612	0.885	0.525	0	2.359	0.276	2.794
G	1.702	3.474	0.947	1.951	0.877	2.416	2.359	0	2.450	3.925
H	2.120	1.149	0.112	0.860	0.856	0.271	0.276	2.450	0	3.677
I	2.601	2.599	2.752	2.598	0.332	1.846	2.322	2.614	2.112	0
L	2.027	2.508	2.680	2.566	1.144	1.790	2.273	2.647	2.059	1.404
K	2.304	1.069	1.250	1.917	1.020	0.895	1.414	2.878	1.144	2.552
M	2.724	1.389	3.471	3.209	0.941	2.385	2.886	3.199	2.656	2.035
F	2.930	2.738	3.170	3.187	0.938	2.261	2.782	3.469	2.528	2.566
P	0.965	2.430	1.876	1.860	0.985	1.470	1.790	1.655	1.677	3.171
S	1.192	2.492	0.699	1.160	0.131	1.443	1.434	0.974	1.491	2.571
T	1.077	2.158	1.112	1.305	0.426	1.145	1.331	1.412	1.285	1.780
W	5.287	2.566	2.943	3.588	0.795	2.562	3.072	4.245	2.797	3.434
Y	3.102	2.449	2.886	3.237	0.662	2.242	2.765	3.765	2.495	2.660
V	1.527	2.536	2.356	2.225	0.759	1.650	2.056	2.076	1.893	0.666

	L	K	M	F	P	S	T	W	Y	V
A	2.151	0.499	2.651	0.972	0.157	0.221	0.858	5.362	3.015	1.915
R	3.078	0.259	3.085	0.944	0.480	0.451	2.102	4.255	3.206	2.553
N	2.595	0.264	2.774	0.823	0.365	0.127	0.854	4.614	2.808	2.210
D	2.306	0.507	3.247	1.207	0.595	0.335	1.305	4.671	2.868	1.891
C	2.004	0.430	1.726	0.825	0.418	0.037	1.602	4.646	2.416	1.337
Q	2.334	0.327	2.479	1.070	0.590	0.311	1.145	4.186	2.596	2.055
E	2.021	0.507	2.964	1.213	0.593	0.489	1.331	5.064	2.699	1.682
G	3.006	0.436	3.821	1.023	0.329	0.140	1.412	6.760	3.520	3.282
H	2.983	0.250	3.435	0.833	0.380	0.308	1.285	3.981	1.297	3.239
I	1.346	1.015	1.730	0.255	0.705	1.029	1.473	4.029	2.657	0.660
L	0	0.996	0.812	0.234	0.711	1.078	1.472	4.509	2.421	0.611
K	1.511	0	1.732	0.977	0.487	0.406	1.466	4.005	2.177	1.524
M	1.165	0.541	0	0.717	0.440	0.530	2.110	3.743	1.910	1.292
F	1.377	1.200	1.910	0	0.856	1.293	2.242	3.137	0.460	2.551
P	2.591	0.433	3.426	0.856	0	0.375	0.572	4.703	2.494	2.550
S	2.610	0.406	2.909	0.922	0.327	0	0.544	4.600	2.826	1.939
T	2.245	0.382	2.617	0.848	0.227	0.045	0	4.923	2.939	1.228
W	2.347	1.178	2.995	0.143	0.921	1.259	2.907	0	1.081	3.474
Y	1.494	1.008	1.959	0.376	0.895	1.227	2.453	3.617	0	2.671
V	1.554	0.900	1.927	0.709	0.586	0.897	0.996	4.708	2.676	0

matches and 0 for mismatches. This matrice emphasize the identities between sequences rather than chemical similarities. The number of true positives found was 2 039 for the control matrix and 4 478 for PAM 250. FITMAT 250 could detect 4 081 true positives, i.e., 91% of the true positives detected by PAM 250, and performed much better than the simple control matrix. With the same procedure, it was found that the performance of FITMAT 250 is 157% of PAM 10, 120% of PAM 50, 97% of PAM 100, and 104% of PAM 450. The sequence alignment performances of the chemically derived FITMAT 250 matrix are rather comparable to those of the phylogenetic PAM 250 matrix and better than the PAM matrix at both close and distant evolutionary distance. This comparison confirms that the chemical information included in FITMAT 250 contributes to improve the alignment performance in contrast to the simple identity matrix.

Extension to noncoded amino acids DISMAT 250 has been extended to 57 noncoded amino acids (20 N-methylated amino acids and 17 various amino acids). A principal components

Table 4. FITMAT 250 20 × 20

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	2	0	0	0	−2	0	−1	3	2	0	0	1	0	−3	3	1	1	−6	−2	1
Arg	−2	6	0	0	0	1	0	−1	3	0	0	2	1	−2	0	0	0	6	−1	−1
Asn	0	2	2	3	2	2	2	1	4	0	0	3	0	−1	1	1	1	0	−2	0
Asp	−1	2	1	4	0	2	2	0	3	−1	−1	0	−1	−5	−1	0	1	1	−4	0
Cys	0	2	−2	−2	12	−1	−2	1	1	3	3	1	3	−1	0	1	0	−5	−1	1
Gln	−1	4	1	2	1	4	3	0	5	0	0	2	0	−4	−1	0	1	3	−3	0
Glu	−1	3	1	2	0	3	4	0	4	0	−1	0	−1	−6	−1	0	1	3	−4	0
Gly	1	0	0	0	0	0	−1	5	1	0	0	1	0	−3	1	1	0	−8	−2	0
His	0	4	1	1	0	3	2	0	6	0	1	3	0	−1	0	0	1	5	0	−1
Ile	0	1	−2	−1	7	0	−1	0	2	5	5	−2	4	5	−3	−1	0	0	4	2
Leu	0	1	−2	−1	−4	0	−1	0	2	4	6	−2	4	6	−3	−2	0	0	4	2
Lys	−1	3	0	0	1	1	0	0	4	2	3	5	2	0	0	0	0	1	−1	1
Met	−1	3	−2	−2	−1	−1	−2	0	1	3	4	0	6	0	0	0	0	−2	0	1
Phe	−1	0	−2	−2	−1	0	−2	−1	1	3	3	−4	4	9	−5	−3	0	6	5	1
Pro	0	1	−1	0	−1	0	0	1	3	2	3	1	2	−1	6	0	1	−3	−1	2
Ser	0	1	1	1	10	1	1	2	3	0	0	1	0	−2	1	2	2	−3	−2	0
Thr	0	2	0	1	5	1	1	2	4	0	0	2	0	−1	2	1	3	−1	−1	1
Trp	−3	1	−2	−3	0	−1	−2	−3	0	1	1	−4	2	7	−6	−2	−1	17	5	0
Tyr	−2	1	−1	−1	2	0	−1	−1	2	2	2	−3	2	4	−5	−2	0	10	10	0
Val	0	1	−1	0	1	0	−1	1	2	3	4	−2	3	0	−1	−1	0	−2	0	4

analysis has been performed over the 77-amino acid databank. The distribution of the amino acids into the two first principal components PC1/PC2 plane (Figure 1) shows a clustering of aromatic, apolar, basic, thiol, and sulfhydryl residues. The substitution coefficients for the following 57 noncoded amino acids of the extended databank are given (Table 5): aminoheptanoic acid (Aha), aminooctanoic acid (Aoa), aminopentanoic acid (Apa), 2-aminobutyric acid (Abu), α-aminoisobutyric acid

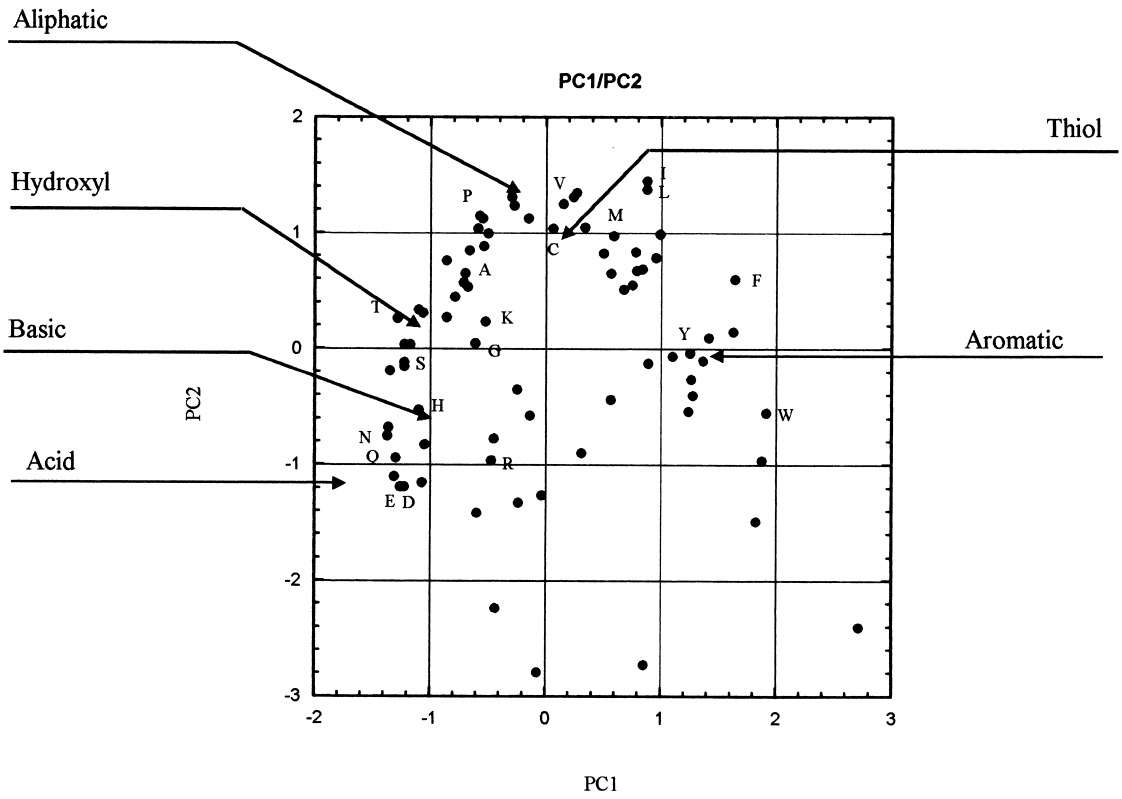

Figure 1. Principal components analysis projection of the amino acids databank mutability shapes (PC1, principal components 1; PC2, principal components 2).

Table 5. Extension of DISMAT 250 to 57 Noncoded Amino Acids

	A	R	N	D	C	Q	E	G	H	I
NMet-Ala	2.220	2.713	1.394	1.911	0.726	1.699	1.946	1.361	1.875	2.056
NMet-Arg	3.885	0.699	0.498	2.629	0.811	1.668	2.019	4.006	1.775	3.902
NMet-Asn	2.305	1.137	0.301	1.065	0.752	0.139	0.609	2.321	0.386	2.456
NMet-Asp	2.213	1.229	0.176	0.779	0.888	0.371	0.177	2.439	0.104	2.167
NMet-Cys	3.504	1.905	2.722	3.237	1.063	2.556	3.014	2.886	2.818	1.50
NMet-Gln	4.614	0.765	0.486	1.528	1.278	0.486	0.983	2.753	0.708	2.660
NMet-Glu	3.285	0.962	0.010	1.290	0.851	0.428	0.681	2.838	0.440	2.430
NMet-Gly	1.554	3.006	0.999	1.767	0.752	1.953	2.020	0.661	2.046	3.514
NMet-His	5.369	0.217	0.218	1.783	0.730	0.839	1.174	3.237	0.927	3.538
NMet-Ile	3.594	2.766	2.329	3.083	0.451	2.196	2.710	3.272	2.466	0.70
NMet-Leu	2.899	2.672	2.275	3.10	0.556	2.181	2.701	3.377	2.449	1.368
NMet-Lys	4.001	1.303	1.223	2.559	1.004	1.519	2.015	3.512	1.739	2.460
NMet-Met	4.676	1.540	2.859	3.676	0.770	2.774	3.293	3.781	3.043	1.801
NMet-Phe	3.662	3.097	2.637	3.655	0.817	2.691	3.215	3.977	2.952	2.824
NMet-Pro	3.167	2.283	1.649	2.186	0.699	1.494	1.941	2.276	1.751	3.005
NMet-Ser	1.675	2.064	0.815	1.293	1.175	1.063	1.276	1.512	1.214	2.038
NMet-Thr	3.756	1.696	1.120	1.717	0.510	0.951	1.412	2.230	1.210	1.192
NMet-Trp	5.698	3.125	2.470	4.249	0.806	3.212	3.709	4.949	3.434	3.824
NMet-Tyr	5.769	2.906	2.427	3.844	0.613	2.823	3.336	4.430	3.062	3.041
NMet-Val	4.741	2.570	2.036	2.632	0.278	1.851	2.335	2.697	2.119	0.318
Aha	2.540	2.213	1.893	2.723	0.520	1.790	2.311	3.137	2.058	3.557
Aoa	3.483	2.523	2.184	3.295	1.102	2.301	2.824	3.805	2.555	4.298
Apa	2.087	2.170	1.309	1.798	0.413	1.272	1.643	1.871	1.499	2.814
Abu	0.741	2.706	1.439	1.946	1.338	1.699	1.963	1.425	1.883	2.010
Aib	0.568	2.648	1.243	1.774	1.421	1.628	1.841	1.250	1.788	3.465
Asu	3.218	1.549	0.658	2.431	0.751	1.427	1.831	3.709	1.571	7.644
Cha	3.797	2.916	2.479	3.685	1.031	2.686	3.207	4.161	2.937	6.970
Chg	3.170	2.610	2.241	3.132	0.599	2.189	2.712	3.477	2.455	6.843
Cit	2.953	0.150	0.301	1.950	0.832	0.991	1.342	3.370	1.093	4.011
Dea	5.524	2.734	2.320	3.125	1.365	2.215	2.733	3.369	2.484	6.821
Dip	8.087	2.771	2.013	4.212	0.751	3.173	3.636	5.128	3.365	7.297
F2-Phe	4.113	2.849	2.439	3.543	1.276	2.557	3.081	3.973	2.813	6.925
F3-Phe	4.380	2.849	2.439	3.545	1.216	2.559	3.083	3.977	2.815	6.936
F4-Phe	3.380	2.850	2.439	3.555	0.576	2.567	3.091	3.992	2.823	6.940
Hci	3.602	0.674	0.594	2.541	0.679	1.555	1.934	3.864	1.681	4.510
Hyp	1.426	1.735	0.775	1.284	1.270	0.786	1.107	1.846	0.983	3.346
NO2-Phe	3.782	2.752	2.302	3.751	1.211	2.725	3.235	4.382	2.961	7.061
Nle	2.048	2.550	2.023	2.622	1.062	1.838	2.323	2.698	2.106	6.908
Nva	1.611	2.536	1.731	2.220	1.091	1.648	2.052	2.068	1.891	6.830
OP-Ser	2.382	0.802	0.318	1.333	0.893	0.297	0.779	2.656	0.504	7.573
OP-Tyr	4.887	2.474	1.675	4.083	0.949	3.054	3.492	5.123	3.226	7.559
Orn	1.769	1.345	0.629	1.272	0.341	0.477	0.921	2.189	0.725	7.131
Pen	4.029	1.622	1.309	2.075	0.539	1.152	1.668	2.687	1.422	1.834
PhG	2.649	2.580	2.151	2.866	0.767	1.995	2.505	3.068	2.266	6.799
Sar	1.365	3.028	0.999	1.780	0.960	1.975	2.039	0.640	2.066	3.526
Sta	3.667	1.996	1.742	3.144	0.975	2.109	2.614	3.928	2.338	2.250
Thi	2.764	1.873	1.604	2.451	0.293	1.501	2.024	2.994	1.768	1.373
b-Ala	1.457	3.018	0.831	1.678	0.749	1.964	1.983	0.525	2.033	3.318
bOH-Val	2.652	1.636	0.876	1.431	0.639	0.769	1.173	2.034	1.004	2.140
d-Pro	0.942	2.584	1.403	1.893	0.825	1.594	1.877	1.509	1.786	3.211
diI-Tyr	15.66	5.391	3.980	6.201	1.106	5.168	5.669	6.727	5.393	8.076
e-Ahx	2.135	2.108	1.601	2.203	1.030	1.418	1.898	2.462	1.684	3.164
g-Abu	1.727	2.404	1.016	1.525	1.277	1.390	1.584	1.289	1.537	2.983
gCOO-Glu	3.955	0.419	1.557	1.959	0.915	1.930	1.645	3.908	1.711	7.427
h-Ser	1.099	1.919	0.553	1.030	1.230	0.891	1.031	1.558	1.001	7.092

Table 5. (Continued)

	A	R	N	D	C	Q	E	G	H	I
h-Ser-Lne	2.159	2.498	1.075	1.597	0.738	1.480	1.671	1.240	1.627	2.068
pCl-Phe	3.770	3.238	2.718	3.970	1.139	2.971	3.492	4.406	3.221	7.044
	L	K	M	F	P	S	T	W	Y	V
NMet-Ala	1.837	0.520	2.315	0.990	0.160	0.394	0.623	3.611	2.459	1.460
NMet-Arg	2.929	0.510	2.922	1.139	0.555	0.511	2.613	3.191	3.016	2.719
NMet-Asn	2.146	0.315	2.351	0.871	0.380	0.054	1.020	3.986	2.338	1.888
NMet-Asp	1.821	0.498	2.813	1.220	0.606	0.289	1.321	4.452	2.343	1.569
NMet-Cys	1.723	0.436	1.558	0.801	0.348	0.186	2.010	2.907	1.941	1.045
NMet-Gln	1.963	0.513	2.142	1.077	0.421	0.419	1.378	2.766	2.335	1.854
NMet-Glu	1.650	0.507	2.606	1.211	0.610	0.465	1.570	3.228	2.332	1.587
NMet-Gly	2.216	0.301	3.194	1.022	0.328	0.199	0.840	6.428	2.939	2.851
NMet-His	2.737	0.511	3.203	0.818	0.429	0.493	1.891	1.139	0.542	3.253
NMet-Ile	1.470	0.429	1.607	0.638	0.455	0.435	2.078	3.449	2.434	1.318
NMet-Leu	0.826	0.739	0.747	1.191	0.751	0.755	2.147	3.892	2.069	1.374
NMet-Lys	1.531	0.286	1.626	1.105	0.398	0.393	2.103	2.429	1.912	1.878
NMet-Met	1.125	0.216	0.659	0.815	0.368	0.341	2.652	2.272	1.597	1.495
NMet-Phe	1.692	0.836	1.961	0.141	0.630	0.894	2.750	2.370	0.075	2.906
NMet-Pro	2.353	0.428	3.176	0.882	0.032	0.135	1.060	2.242	2.045	2.513
NMet-Ser	2.208	0.501	2.518	0.964	0.308	0.258	0.099	3.987	2.375	1.446
NMet-Thr	1.731	0.350	2.092	0.852	0.223	0.109	0.838	3.231	2.428	0.874
NMet-Trp	2.812	0.564	3.247	0.555	0.640	0.753	3.611	0.646	1.423	3.975
NMet-Tyr	2.013	0.601	2.205	0.564	0.622	0.746	3.117	0.811	0.604	3.162
NMet-Val	1.345	0.439	1.575	0.801	0.305	0.319	1.533	3.136	2.351	0.663
Aha	1.934	1.017	2.056	0.485	0.874	1.052	1.838	4.968	2.948	1.591
Aoa	2.818	1.209	2.760	0.524	0.970	1.241	2.50	2.874	3.593	2.067
Apa	1.181	0.357	1.834	0.946	0.372	0.252	0.608	6.155	2.367	1.533
Abu	1.786	0.426	2.264	1.008	0.205	0.374	0.649	5.147	2.415	1.424
Aib	1.527	0.486	2.70	1.031	0.302	0.311	0.510	4.736	2.641	2.576
Asu	2.376	0.429	6.535	1.090	0.541	0.434	2.307	5.979	5.216	6.760
Cha	1.902	0.743	6.097	0.916	0.765	0.621	2.879	4.489	4.473	6.826
Chg	1.876	0.550	6.151	0.651	0.430	0.477	2.222	4.582	4.541	6.634
Cit	2.956	0.299	2.987	0.891	0.539	0.458	2.007	4.023	3.210	2.405
Dea	1.511	0.450	6.047	1.174	0.539	0.310	2.153	4.077	4.511	6.614
Dip	3.029	0.975	6.538	0.637	0.881	1.035	3.744	4.870	4.734	7.208
F2-Phe	1.804	0.856	6.081	0.680	0.647	0.811	2.702	4.10	4.470	6.768
F3-Phe	1.659	0.522	6.040	0.843	0.466	0.499	2.706	3.977	4.463	6.769
F4-Phe	1.672	0.330	6.043	0.805	0.469	0.292	2.719	4.950	4.463	6.773
Hci	3.389	0.222	3.512	0.867	0.543	0.444	2.465	3.880	4.126	2.546
Hyp	2.682	0.469	3.475	0.953	0.227	0.463	0.438	3.932	2.117	2.778
NO2-Phe	1.916	0.519	6.111	0.986	0.439	0.458	3.054	4.903	4.510	6.897
Nle	0.068	0.320	5.920	0.999	0.468	0.309	1.529	5.378	4.693	6.504
Nva	0.939	0.343	6.045	1.006	0.442	0.306	0.989	5.355	4.777	6.470
OP-Ser	2.453	0.360	6.420	1.115	0.596	0.348	1.320	5.013	5.022	6.766
OP-Tyr	2.777	0.723	6.449	1.127	0.970	0.802	3.721	4.416	4.922	7.232
Orn	1.712	0.361	6.212	0.962	0.464	0.418	0.803	5.360	4.735	6.632
Pen	1.614	0.360	1.951	1.048	0.273	0.254	1.304	2.0	1.235	1.584
PhG	1.724	0.947	6.135	0.475	0.693	0.840	1.857	4.855	4.593	6.554
Sar	2.234	0.343	3.209	1.023	0.334	0.229	0.863	6.281	2.963	2.861
Sta	1.506	0.731	2.011	0.542	0.542	0.616	2.551	4.111	3.042	2.149
Thi	1.608	0.182	1.761	0.954	0.465	0.170	1.644	3.222	1.461	1.336
b-Ala	2.40	0.426	3.108	0.944	0.196	0.203	0.90	6.402	2.992	2.613
bOH-Val	1.957	0.431	2.325	1.048	0.289	0.258	0.621	2.934	1.703	1.670
d-Pro	2.642	0.408	3.479	0.840	0.022	0.365	0.589	4.319	2.629	2.574

Table 5. (Continued)

	L	K	M	F	P	S	T	W	Y	V
diI-Tyr	4.495	0.663	6.965	0.791	0.659	0.610	5.469	11.39	5.062	8.158
e-Ahx	1.432	0.404	1.818	0.874	0.415	0.403	1.179	5.543	2.664	1.367
g-Abu	1.633	0.352	2.327	0.947	0.192	0.211	0.253	6.219	2.524	1.970
gCOO-Glu	2.764	0.239	7.214	1.058	0.461	0.192	2.972	4.662	4.520	7.010
h-Ser	2.086	0.407	6.350	1.048	0.353	0.303	0.302	5.973	4.867	6.646
h-Ser-Lne	2.958	0.398	3.232	1.022	0.405	0.229	0.342	4.076	3.264	1.086
pCl-Phe	2.065	0.492	6.105	1.020	0.590	0.444	3.145	4.966	4.483	6.921

(Aib), α-aminosuberic acid (Asu), β-cyclohexylalanine (Cha), cyclohexylglycine (Chg), citrulline (Cit), 2-fluorophenylalanine (F2-Phe), 3-fluorophenylalanine (F3-Phe), 4-fluorophenylalanine (F4-Phe), homocitrulline (Hci), (Hyp), 4-nitrophenylalanine (NO2-Phe), norleucine (Nle), norvaline (Nva), phosphoserine (OP-Ser), phosphotyrosine (OP-Tyr), ornithine (Orn), penicillamine (Pen), phenylglycine (Phg), sarcosine (Sar), statine (Sta), β-(2-Thienyl)alanine (Thi), β-alanine (b-Ala), β-hydroxyvaline (bOH-Val), 3,4-dehydroproline (d-Pro), 3,5-diiodotyrosine (diI-Tyr), 6-aminocaproic acid (e-Ahx), 4-aminobutyric acid (g-Abu), γ-carboxyglutamic acid (gCOO-Glu), homoserine (h-Ser), homoserine-lactone (hSer-Lne), and 4-chlorophenylalanine (pCl-Phe).

Use of the automated sequence mutation strategy The choice of an amino acid substitution can be quantitatively guided by reference to a distance matrix, especially DISMAT 250. In an appropriate descriptor space, one can find a maximum distance compatible with biological activity. This cutoff score could be chosen by the user or automatically computed from a mathematical description of the quantitative structure–activity relationship in a test series. The automated sequence mutation (ASM) program provides the list of all amino acids on the condition that the matrix score remain lower than the cutoff value.

Analysis of Peptide Quantitative Structure–Function Relationships

Linear regression analysis Hitherto, only linear methods such as PLS[18] or linear regression[32–34] have been routinely used. Here, stepwise multiple regression was used to analyze simultaneously the influence of n properties at p positions. The interest of the stepwise mode is to exclude from the model every position that is not significant in the analysis and to detect nonlinearities in the response curve. If a single position is varied, the model can exhibit a linear or linearizable relationship between the biological activity and the distance score along this position. If this is not the case, the activity will be assumed to be sensitive to another set of physicochemical properties, or will be intrinsically nonlinear, or the position will have no effect on the activity. The influence of mutation on dynamic trajectories or conformation is not linear free energy related and is not expected to induce a linear behavior in activity. The correlation coefficients between activity data and the distance scores extracted from each of the 140 distance matrices are systematically checked. The predictive power of the model is evaluated by a cross-validation procedure: The table rows are sorted by Y values and then a group of data is withheld from the calculations in a systematic way. The remaining data in X is then used to produce a new model for Y. Predicted values are compared with the exact values for each group of rows that have been withheld and $r(CV)^2$, the cross-validated equivalent of the determination coefficient r^2, is supplied with the regression equation.

Detection of distance constraints along position axis by distance mapping analysis If regression methods fail to detect linearity, distance mapping provides a good alternative. This method is an extension to distances of the nonlinear detection of constraints algorithm, called variable mapping, which gives successful results in our laboratory.[35] Distance mapping searches for a significant vicinity between the reference compound and the active compounds along the axes corresponding to the different key positions. If the reference compound is at the center of the activity range, active and inactive compounds will be perfectly discriminated.

Since the leader compound is not supposed to be at the center of the activity range, a close compound can be either active or inactive while a more distant compound can be an active one. When using distance as a descriptor, the data in the activity range are embedded. In order to circumscribe the activity range, the more distantly active compounds are spotted providing a cutoff, or maximal distance λ compatible with activity. The more analogs one obtains, the more the set of descriptor limit values will be refined, even if only one active analog is sufficient to define an activity range. Thus, for a new

Table 6. Biological Data for 12 Analogs of ET-1 Varied in Position 21

Analog	ln(activity)	ln(affinity)
1. ET-1	4 605	4 605
2. Tyr^{21}	2 862	3 466
3. Phe^{21}	2 477	1 609
4. His^{21}	0 182	1 792
5. Gly^{21}	−1 050	−1 022
6. Ser^{21}	−1 897	2 197
7. Ala^{21}	−2 303	−1 609
8. Lys^{21}	−2 408	−2 813
9. Ile^{21}	−3 912	1 609
10. Pro^{21}	−3 912	−3 912
11. Glu^{21}	−3 912	−1 139
12. Gln^{21}	−3 912	−6 215

Data taken from Koshi et al.[38]

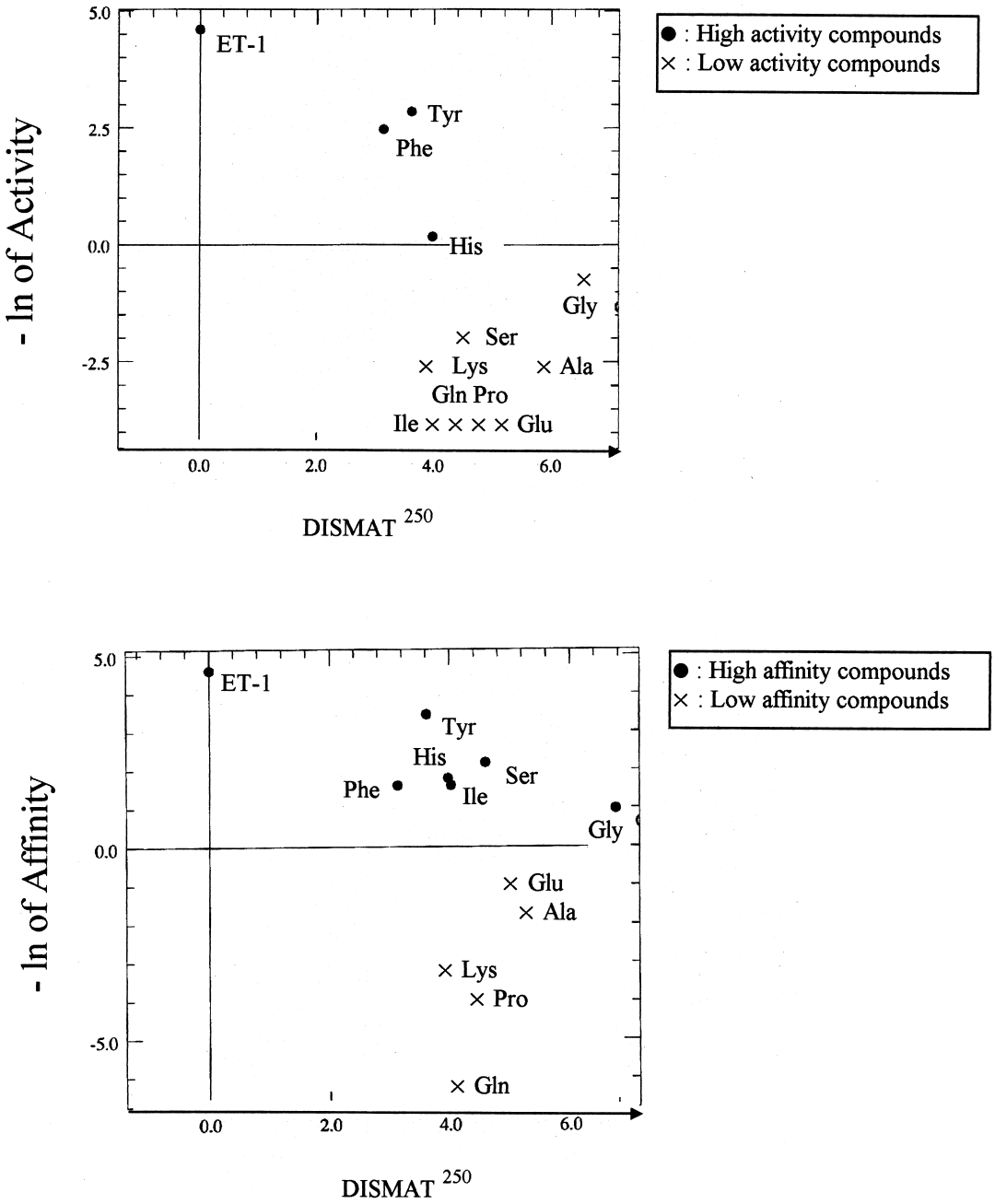

*Figure 2. Plot of activity (*top*) or affinity (*bottom*) versus DISMAT 250 at position 21 of ET-1.*

given peptide, distance mapping can predict whether or not it falls within the activity range.

The extensibility $E = \lambda/D_{max}$, where D_{max} is the maximal observed distance in the training set, measures the acceptability at each position for chemical change. A key position is characterized by a small extensibility value. A simple examination of the graphical representation gives a diagnosis of the qualitative nonlinear dependencies between the activities and distances from the reference compound.

Cluster significance analysis is useful to evaluate if the clustering of active compounds in a narrow range of distances has arisen merely by chance.[36] In fact, if a position were not related to the biological activity, via a particular property, the active compounds would be scattered randomly about the graph. The cluster significance analysis algorithm groups a set of points that consists of similar members, based on their distances along the position axes. Mean square distance is defined by

$$\mathrm{MSD} = \sum_{j=1}^{N} d^2(j, j')/N$$

for N compounds, j separated by a Euclidian distance d in the n-space of positions. The algorithm compares the validity of the a priori classification by testing all possible combinations of individual points and comparing the MSD of clusters. The

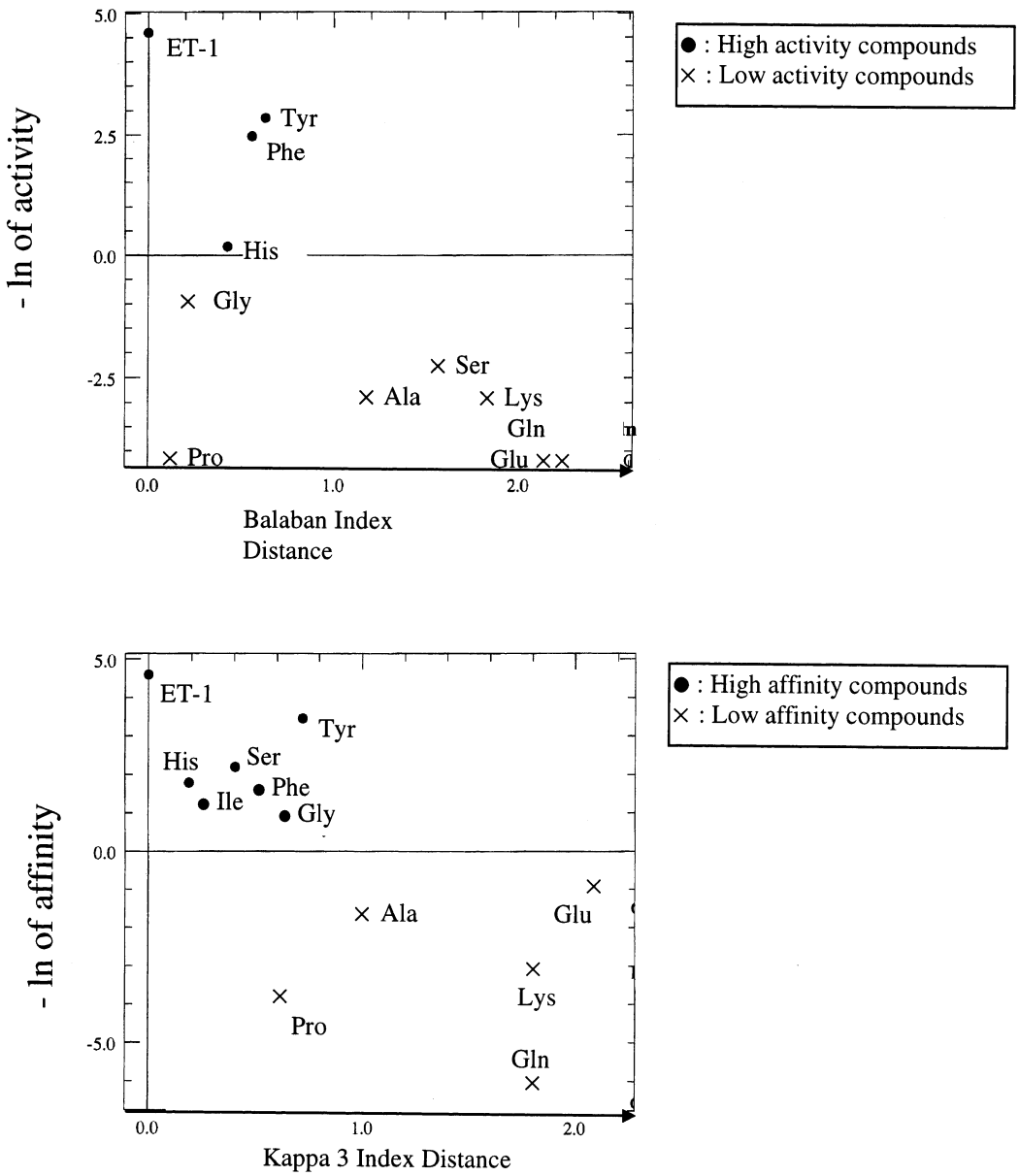

*Figure 3. Plot of activity versus Balaban index distance (*top*) and affinity versus Kappa3 distance (*bottom*) at position 21 of ET-1.*

cluster significance is the probability that the proposed classification is more valid than that obtained by chance alone.

MATERIALS/MODELLING

All calculations of descriptors and statistical analysis were performed with TSAR, ASP, and VAMP[24] software on a Silicon Graphics Indy workstation. The 20 natural amino acids were built in MAD.[24] All structures were optimized by a Newton–Raphson procedure, using the COSMIC force field. The initial conformation corresponds to the mean structure as found in the PDB. Side chains are represented as they exist under physiological conditions at pH 7: Asp and Glu are nonprotonated and Arg, Lys, and His are protonated. Alignments of sequences were performed with BLAST V.1.4 from the NCBI on a Silicon Graphics Indy workstation and performances of FITMAT 250 were evaluated using a program from Henikoff and Henikoff,[30] available by ftp.

RESULTS

Monopositional Analysis of 11 Analogs of endothelin 1 Varied in Position 21

Endothelin 1 (ET-1) is a potent vasoconstrictor isolated from cultured porcine endothelial cells. This peptide contains 21 amino acids arranged in a unique bicyclic motif formed by disulfide bridges. The C-terminal hexapeptide has been showed to discriminate between different endothelin receptors[37] and the C-terminal Trp in position 21 seems to be crucial for biological activity. Koshi et al.[38] synthesized 11 analogs varied at position 21 and examined their

Table 7. Biological Data for 21 C-Terminal ET-1 Hexapeptides

Analog	ln(ET-A)	ln(ET-B)
1. D-Dip[16]-LDIIW	−3.129	−2.813
2. D-Dip[16]-EDIIW	−3.689	−2.957
3. D-Dip[16]-RDIIW	−5.521	−4.605
4. D-Dip[16]-LEIIW	−3.863	−3.693
5. D-Dip[16]-LKIIW	−0.734	−3.411
6. D-Dip[16]-LFIIW	−0.545	−2.813
7. D-Dip[16]-LYIIW	−0.693	−2.526
8. D-Dip[16]-LAIIW	−3.352	−4.135
9. D-Dip[16]-LDEIW	0.0	1.792
10. D-Dip[16]-LDKIW	2.303	1.308
11. D-Dip[16]-LDFIW	−1.609	−2.882
12. D-Dip[16]-LDYIW	−2.688	−4.423
13. D-Dip[16]-LDAIW	−2.303	−1.109
14. D-Dip[16]-LDVIW	−4.075	−2.976
15. D-Dip[16]-LDIEW	2.303	ND
16. D-Dip[16]-LDIKW	2.303	ND
17. D-Dip[16]-LDIFW	1.386	ND
18. D-Dip[16]-LDIVW	−1.966	ND
19. D-Dip[16]-LDILW	−1.470	ND
20. D-Dip[16]-LDIIF	1.609	ND
21. D-Dip[16]-LDIIY	0.788	ND

Data taken from Doherty et al.[39]

agonistic vasoconstrictor activity on rat thoracic aortic strips, and receptor binding activity on rat brain membrane fractions (Table 6). We applied DISMAT 250 scores to analyze the results of Koshi. The reference compound was ET-1 with Trp at position 21. The analysis of activity and affinity does not exhibit any linear relationship (Figure 2) but it should be noted that 92% of the compounds were well classified using DISMAT 250 distance criteria in a discriminant analysis. From a distance mapping point of view, there is no cutoff, indicating that phylogenetic distance does not really influence the activity or the affinity at this position, even if the presence of an aromatic side chain always leads to the most potent compounds. Either aromatic substitutions such as [Tyr[21]]-; [Phe[21]]-; and [His[21]]ET-1, or polar substitutions such as [Ser[21]]ET-1, or apolar substitutions such as [Ile[21]]- and [Gly[21]]ET-1 could lead to active compounds. The chemical nature at this position does not really influence the affinity. This does not suggest a direct interaction between the Trp[21] side chain and the ET receptor. Extensibility, as well as correlation coefficients between biological data and the distance scores extracted from each of the 140 distance matrices, have been calculated. No linear relationships have been detected, but we have found significant topological distance constraints for both activity and affinity, using, respectively, Balaban index ($p = 0.99$) and Kappa 3 criteria ($p = 0.97$) (Figure 3). The Balaban index is defined as the average distance sum connectivity and characterizes the degree of linearity of the side chain: the aromatic residues, exhibiting a low value of the Balaban index distance with respect to Trp, conserves the biological activity. The Kappa 3 connectivity index indicates the degree of branching at the center of the side chain. This topological feature is the more accurate among the 140 matrix descriptors in predicting the influence of mutations at position 21 on the affinity.

Table 8. Distance Mapping Results for C-terminal ET-1 Hexapeptides at Position 17 to 21

Position	Range	Extensibility (%)	Significance
	1. ET_A affinity		
17	0–2.365	100	0.00
18	0–0.859	31	0.92
19	0–1.55	64	0.80
20	0–0.553	23	0.93
21	0–0	0	0.68
	2. ET_B affinity		
17	0–2.365	100	0.00
18	0–2.79	100	0.51
19	0–0.923	38	0.99
20	ND	ND	ND
21	ND	ND	ND

Abbreviation: ND, Not determined.

Multipositional Analysis of 21 Analogs of C-Terminal ET-1 Hexapeptides

We have analyzed 21 C-terminal endothelin hexapeptide antagonists from literature SAR data[39–41] (Table 7). The peptide leader is D-Dip[16]-RDIIW and analogs are varied in positions 17 to 21. The linear analysis of positions 18 and 19 leads to significant cross-validated linear relationships between both ET-A and ET-B receptor affinities and DISMAT 250 scores (Figure 4). Moreover, ET-A affinity has been found to decrease exponentially with DISMAT 250 scores at position 20.

Using distance mapping, only position 19 leads to a significant clustering of high-affinity compounds. The 6D distance mapping plot is projected onto the PC1/PC2 plane (Figure 5). The univariate nature of the training set leads to a characteristic representation of the compounds aligned along their respective varied positions, split in distinct directions. All the active compounds are clustered together in a polygon of high affinity. Distance mapping results (Table 8) show that every substitutions in position 19 should correspond to a DISMAT 250 coefficient smaller than 0.923. We have previously pointed out the potential importance of position 19 for the binding of the peptide. In view of the design of an antagonist of clinical use, it is a priority to define all the mutations consistent with an increase of the affinity. The ASM strategy allows listing of all the suitable residues at position 19 within the extended amino acid databank. In this case, the 21 following residues were found among our 77 residue databank to have a DISMAT score lower than 0.923: Cys, Ile, Leu, Met, Phe, Val, Aha, Apa, Chg, Cit, NMet-Cys, NMet-Ile, NMet-Leu, NMet-Pro, NMet-Val, Nle, Nva, Pen, Phg, DPro, and e-Ahx.

Multipositional Analysis of 21 Analogs of Oxytocin Varied in Positions 2, 3, and 8

Oxytocin is a neurohypophyseal nonapeptide with an intramolecular disulfide linkage between cysteines 1 and 6. In vivo

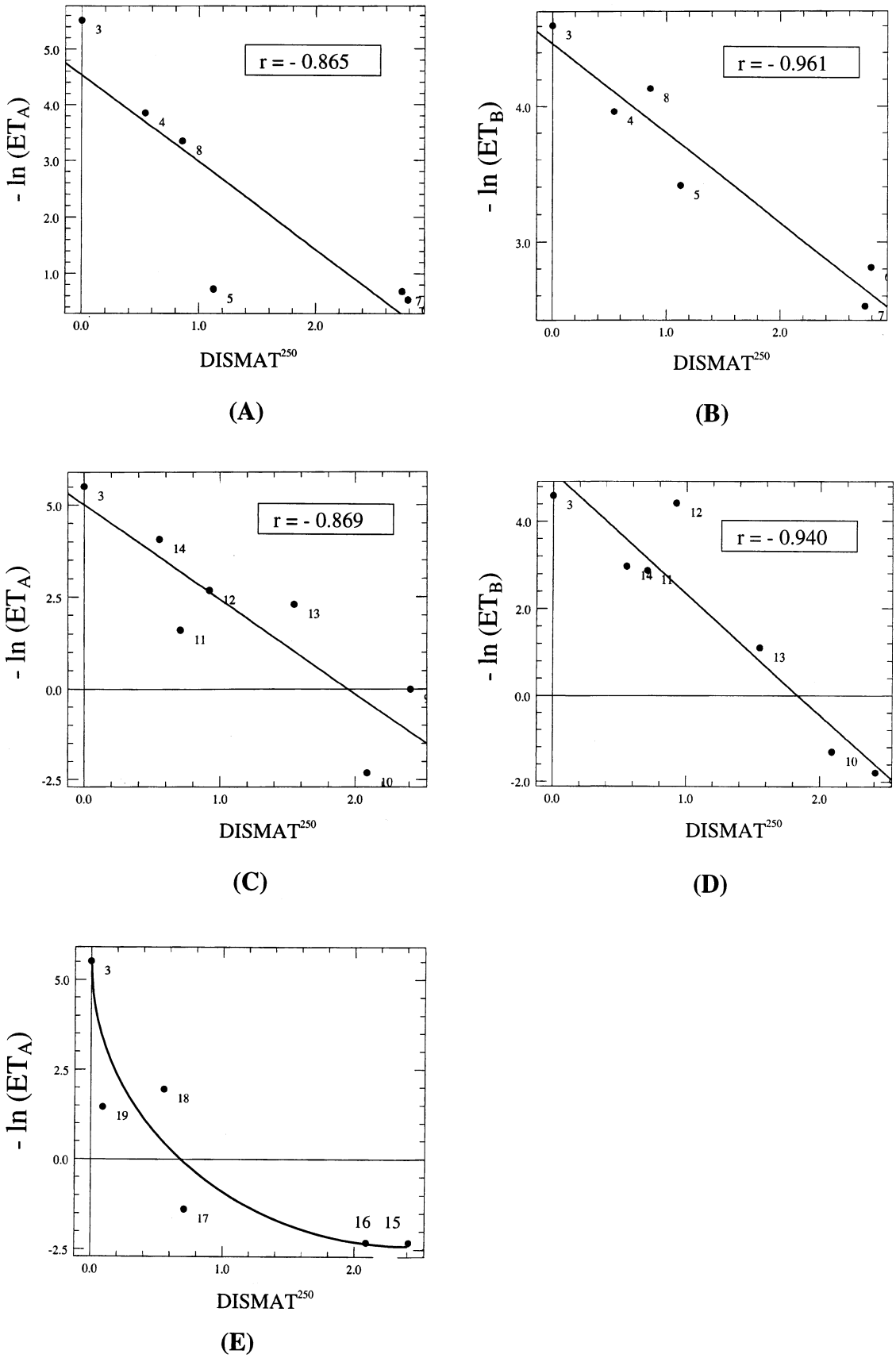

*Figure 4. Relationships between ET_A or ET_B receptor affinities and DISMAT^{250} scores at positions 18 (***A** *and* **B***), 19 (***C** *and* **D***), and 20 (***E***). The reference compound is* **3**.

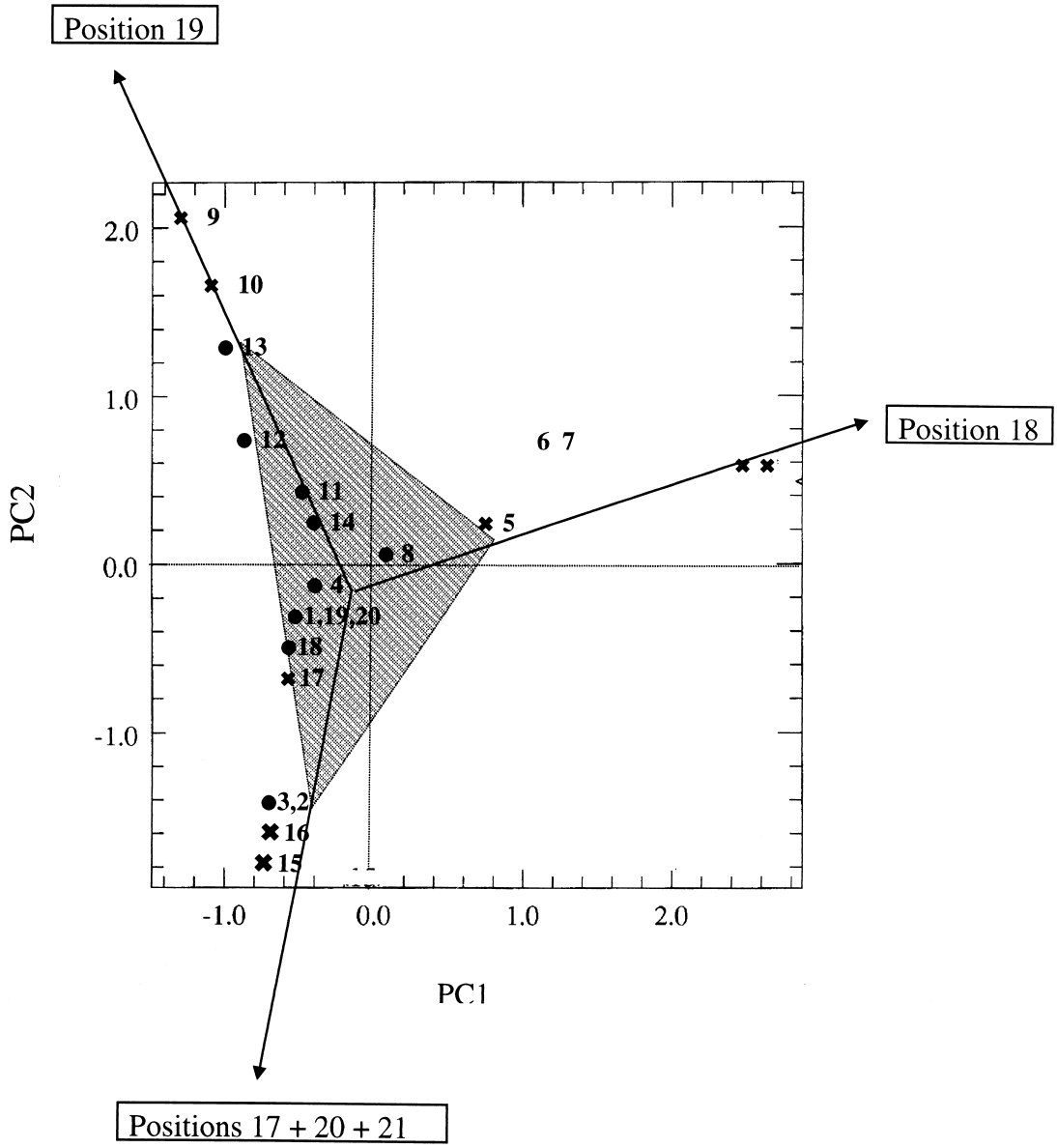

Figure 5. Projection of the 5D distance mapping plot onto the PC1/PC2 plane. The active compounds (shaded region) are clustered in the activity range. The reference compound is **1**.

oxytocic activities on rat uterus were compiled by Sneath[1] (Table 9) and 21 analogs varied in positions 2, 3, and 8 were analyzed with DISMAT 250. Oxytocin was the reference compound. Figure 6 shows that the distribution of active compounds is restricted to a narrow area around the reference compound. Moreover, we have found a good linear relationship between the oxytocic activity (**Y1**) and the DISMAT 250 distances along positions 2 (**XI**), 3 (**X2**), and 8 (**X3**) (Figure 7).

$$\mathbf{Y1} = -3.640\mathbf{X1} - 2.367\mathbf{X2} - 0.388\mathbf{X3} + 5.423$$

$$r = 0.965 \qquad r^2 = 0.931 \qquad r(\mathrm{CV})^2 = 0.907 \qquad n = 21$$

A similar analysis was made with pressive activity, evaluated from rat blood pressure measurements.[1] Vasopressin (ADH) was used as the reference. No significant linear relationships were obtained with the pressive activity, regardless of the position. The 3D distance mapping plot shows a clustering of active compounds around the reference only along two axes corresponding to position 2 and 3. Position 8 seems to be insensitive to chemical changes in the pressive activities and has a moderate effect on oxytocic activity. In both cases, position 8 exhibits an extensibility of 100%. Embedded data indicate that active and inactive compounds are nonlinearly separable.

Results from distance mapping for pressive activity (Table 10) show that every substitution in position 2 should correspond to a DISMAT 250 coefficient smaller than 0.376, but the clustering of active compounds is not really significant (p = 0.612). The only suitable residues at position 2 in the extended database are Tyr, Phe, and NMet-Trp. In position 3, substitutions should correspond to a DISMAT 250 coefficient smaller than 0.365, with a high significance (p = 0.995). ASM suitable residues for position 3 in the extended amino acid databank are Ile, Leu, Met, Val, NMet-Pro, NMet-Val, norleucine, norval-

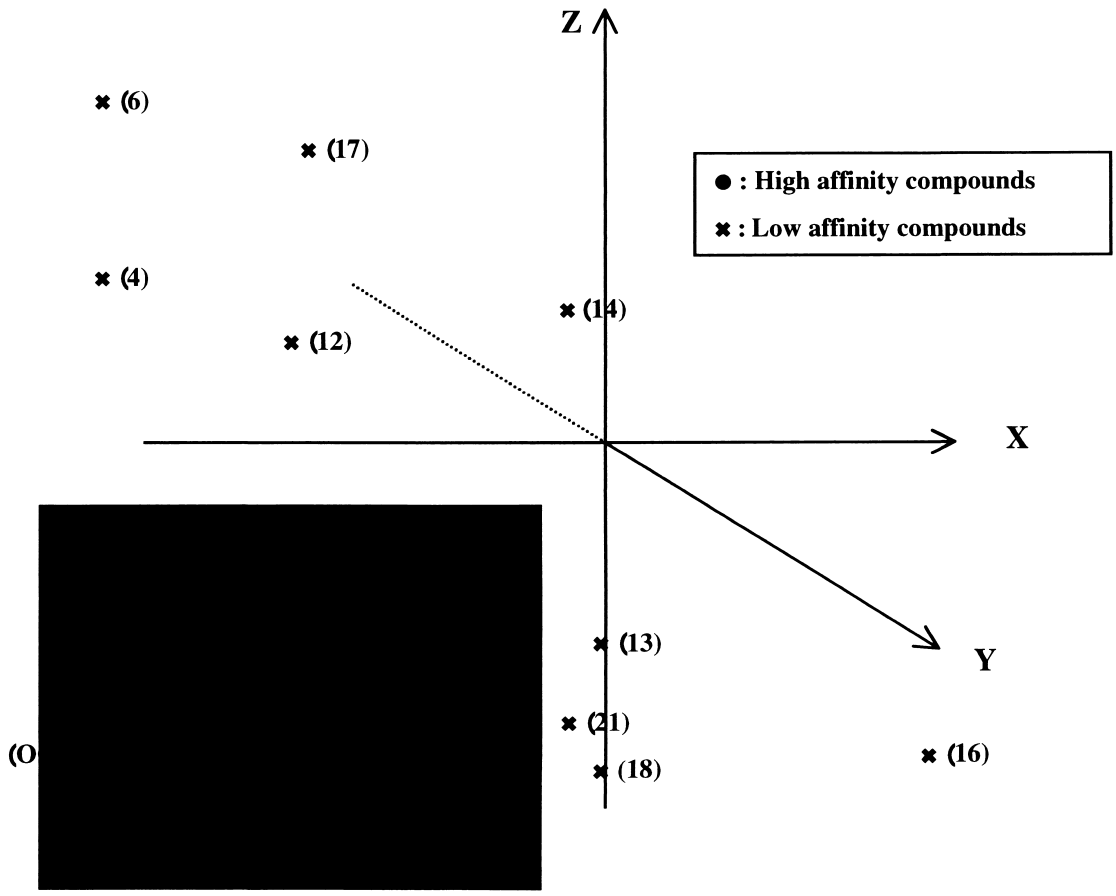

Figure 6. 3D distance mapping plot for oxytocin varied at positions 2 (x axis), 3 (z axis), and 8 (y axis). The activity range lies in the shaded area.

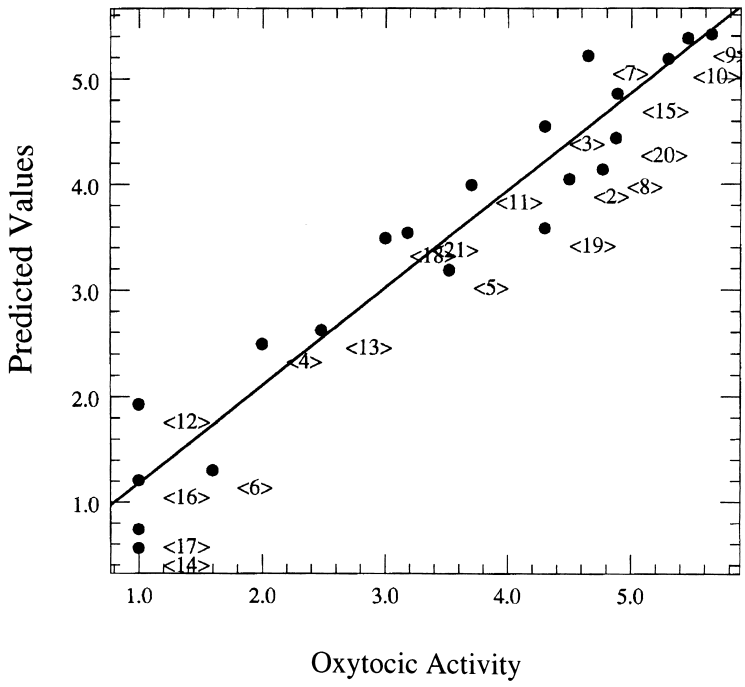

Figure 7. Plot of observed versus calculated oxytocic activity for the training set 1–21.

Table 9. Biological Data for 21 Analogs of Oxytocine Varied at Positions 2, 3, and 8

Analog	Oxytocic activity	Pressive activity
1. YIL (= **OCY**)	5.65	3.70
2. FIL	4.50	2.60
3. YFL	4.30	3.48
4. YYL	2.00	3.00
5. FFL	3.52	2.95
6. YWL	1.60	—
7. YLL	4.65	2.48
8. YVL	4.77	2.30
9. YII	5.46	3.80
10. YIV	5.30	3.95
11. YFK (= **ADH**)	3.70	5.43
12. YYK	1.00	3.20
13. FFK	2.48	4.74
14. FYK	<1.00	2.15
15. YIK	4.89	5.11
16. SIK	<1.00	2.00
17. YWK	<1.00	1.85
18. FIK	3.00	4.51
19. YFR	4.30	5.60
20. YIR	4.88	5.10
21. YFH	3.18	3.18

(Data taken from Sneath.)[1]

ine, phenylglycine, and e-Ahx. All these substitutions are compatible with retention of the pressive activity and lead to a list of potentially active peptides.

CONCLUSION

Substitution matrices have demonstrated their ability to detect linear relationships between the biological activities and affinities of certain peptides. In order to apply this feature to peptide sequence mutation, we calculated a physicochemical matrix fitted on PAM 250 mutability shapes. Extension of this matrix to 57 noncoded amino acids currently used in pharmacomodulation studies leads to a simple and convenient automated tool for peptide or protein sequence mutation. The automated sequence mutation (ASM) strategy is an attempt to generate rules of exchangeability of amino acid side chains in proteins or peptides. It can be useful to design analog test series for quantitative structure–activity or structure–function relationships and to detect relevant properties at different positions of a given sequence. In some models, the observed deviations from linearity can be interpreted as (1) nonlinear perturbations induced by residue replacements, in relation to side chain–side chain interactions, or (2) deviations from the reference molecular dynamics trajectory. Distance mapping analysis is an alternative method for building decision rules about sequence modulation in cases of such nonlinear behavior. A mechanistic interpretation of the model can be deduced from the identification of the most relevant properties at a given position. It should be noted that the matrices developed in this study do not carry any information about the inversion of configuration L/D, the secondary and the three-dimensional structures, or the chemical environment of the target residues. Clearly, these aspects of peptide design cannot be neglected, especially with nonconstrained peptides, and it will be necessary to carry out a systematic evaluation of the impact of the mutation on the molecular conformation. Secondary structure prediction algorithms are available on the web and a graphical program for molecular dynamics trajectory analysis has been constructed in our laboratory. For a particular task, especially when the chemical environment is supposed to be taken into account, distance matrices can even be calculated between different clusters of residues in a window of variable width.

The ASM strategy is proposed as a quantitative or semiquantitative method designed to analyze the relationship between sequence and biological activity in peptides. Searching for amino acid substitutions that satisfy both local and global constraints is a challenge for drug design today. Application of the distance-based approach described here should be helpful for the design of active peptides whose fundamental features such as conformational restriction, oral bioavailability, and cellular targetting have been optimized and that satisfy formulation requirements. Further development for protein engineering puposes will be particularly interesting to provide a mechanistic understanding of directed mutagenesis results.

Table 10. Distance Mapping Results for Oxytocine at Positions 2, 3, and 8

Position	Range	Extensibility (%)	Significance
	1. Oxytocic activity		
2	0–0.376	37.6	0.922
3	0–0.539	30.8	0.998
8	0–2.508	100	0.012
	2. Pressive activity		
2	0–0.376	37.6	0.612
3	0–0.365	20.8	0.995
8	0–2.508	100	0.421

ACKNOWLEDGMENTS

This study was supported by a grant from the CNRS (ACC SV13 INFOBIOSUD). The authors acknowledge a grant of computer time from the CNUSC (F). Thanks are due to Professor A.R. Rees and C. Royer for helpful comments.

REFERENCES

1 Sneath, P.H.A. Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* 1966, **12,** 157–195

2 Miyata, T., Miyazawa, S., and Yasanuga, T. Two types of amino-acid substitutions in protein evolution. *J. Mol. Evol.* 1979, **12,** 219–236

3 Glen, R.C., and Payne, A.W.R. A genetic algorithm for the automated generation of molecules within constraints. *J-CAMD* 1995, **9,** 181–202
4 Kinnear, K.E. *Advances in Genetic Programming* MIT Press, Cambridge, 1994, p. 518
5 Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. *Atlas of Protein Sequence and Structures* (Dayhoff, M.O., ed.). National Biomedical Research Foundation, Washington, D.C., 1979, p. 345
6 Henikoff, S., and Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* 1992, **89,** 10915–10919
7 Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* 1974, **185,** 862–864
8 Epstein, C. Non-randomness of amino acid changes in the evolution of homologous proteins. *Nature (London)* 1967, **215,** 355–359
9 Fauchère, J.L., Quarandon, P., and Kaetterer, L. Estimating and representing hydrophobicity potential. *J. Mol. Graphics* 1988, **6,** 203–206
10 Rao, J.K.M. New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *Int. J. Peptide Protein Res.* 1987, **29,** 276–281
11 Goodman, M., and Moore, G.W. Use of Chou–Fasman amino acid conformational parameters to analyze the organization of the genetic code and to construct protein genealogies. *J. Mol. Evol.* 1977, **10,** 7–47
12 Ptitsyn, O.B., and Finkelstein, A.V. Similarities of protein topologies: Evolutionary divergence, functional convergence or principles of folding. *Q. Rev. Biophys.* 1980, **13,** 339–386
13 Richardson, J.S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 1981, **34,** 167–339
14 Risler, J.L., Delorme, M.O., Delacroix, H., and Henaut, A. Amino acid substitutions in structurally related proteins: A pattern recognition approach. *J. Mol. Biol.* 1988, **204,** 1019–1029
15 Schultz, G.E., and Schirmer, R.H. *The Principles of Protein Structure* Springer-Verlag, New York, 1979
16 Ladunga, I., and Smith, R.F. Amino acid substitutions preserves protein folding by conserving steric and hydrophobicity properties. *Protein Eng.* 1997, **10,** 187–196
17 Jonsson, J., Eriksson, L., Hellberg, S., Sjostrom, M., and Wold, S. Multivariate parametrization of 55 coded and non coded amino acids. *Quant. Struct. Act. Relat.* 1989, **8,** 204–209
18 Hellberg, S., Sjostrom, M., Skagerberg, B., and Wold, S. Peptide quantitative structure–activity relationships, a multivariate approach. *J. Med. Chem.* 1987, **30,** 1126–1135
19 Bogardt, R.A., Jones, B.N., Dwulet, F.E., Garner, W.H., Lehman, L.D., and Gurd, F.R.N. Evolution of the amino acid substitution in the mammalian myoglobin gene. *J. Mol. Evol.* 1980, **15,** 197–218
20 Creighton, T.E. Evolutionary and genetic origins of protein sequences. In: *Proteins: Structure and Molecular Properties*, 2nd Ed. W.H. Freeman and Company, New York, 1993, pp. 105–137
21 Vishwanadhan, V.N., Ghose, A.K., Revankar, G.R., and Robins, R.K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure–activity relationships. 4. Additional parameters for hydrophobic and dispersion interaction. *J. Chem. Inf. Comput. Sci.* 1989, **29,** 163
22 Hall, L.H., and Kier, L.B. *Reviews in Computational Chemistry* (Lipkowitz, K. and Boyd, D., eds.). Indiana University-Purdue University at Indianapolis, Indiana, 1992, p. 367
23 Balaban, A.T. Highly discriminating distance based topological index. *Chem. Phys. Lett.* 1982, **89,** 399–404
24 Oxford Molecular, Ltd. Magdalen Centre, Oxford Science Park, Stanford-on-Thames, Oxford OX4 4GA, England
25 Meyer, A.Y., and Richards, W.G. Similarity of molecular shapes. *J-CAMD* 1991, **5,** 427–440
26 George, D.G., Barker, W.C., and Hunt, L.T. Mutation data matrix and its uses. *Methods Enzymol.* 1990, **183,** 333–351
27 French, S., and Robson, B. What is a conservative substitution? *J. Mol. Evol.* 1983, **19,** 171–175
28 Leunissen, J.A.M., and De Jong, W.W. Phylogenetic trees constructed from hydrophobicity values of protein sequences. *J. Theor. Biol.* 1986, **119,** 189–196
29 Altschul, S.F. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 1991, **219,** 555–565
30 Henikoff, S., and Henikoff, J.G. Performance evaluation of amino acid substitution matrices. *Protein Struct. Funct. Genet.* 1993, **17,** 49–61
31 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* 1990, **215,** 403–410
32 Eroshkin, A.M., Zhilkin, P.A., and Fomin, V.I. Algorithm and computer program Pro-Anal for analysis of relationship between structure and activity in a family of proteins or peptides. *CABIOS* 1993, **9,** 491–497
33 Nomizu, M., Iwaki, T., Yamashita, T., Inagaki, Y., Asano, K., Akamatsu, M., and Fujita, T. Quantitative structure–activity study of elastase substrates and inhibitors. *Int. J. Peptide Protein Res.* 1993, **42,** 216–226
34 Pliska, V., and Heiniger, J. Structural requirements of the oxytocin receptor in rat uterus. *Int. J. Peptide Protein Res.* 1988, **31,** 520–536
35 Grassy, G., Trappe, P., Bompart, J., Calas, B., and Auzou, G. Variable mapping of structure–activity relationships: Application to 17-spironolactone derivatives with mineralocorticoid activity. *J. Mol. Graphics* 1995, **13,** 356–367
36 MacFarland, J.W., and Gans, D.J. On the significance of clusters in the graphical display of structure–activity data. *J. Med. Chem.* 1986, **29,** 505–514
37 Maggi, C.A., Giulani, S., Patacchini, R., Santicioli, P., Rovero, P., Giachetti, A., and Meli, A. The C-terminal hexapeptide, endothelin-(16–21), discriminates between different endothelin receptors. *Eur. J. Pharmacol.* 1989, **166,** 121–122
38 Koshi, T., Suzuki, C., Arai, K., Mizoguchi, T., Torii, T., Hirata, M., Ohkuchi, M., and Okabi, T. Syntheses and biological activities of ET-1 analogs. *Chem. Pharm. Bull.* 1991, **39,** 3061–3063
39 Doherty, A.M., Cody, W.L., DePue, P.L., He, J.X., Waite, L.A., Leonard, D.M., Leitz, N.L., Dudley, D.T., Rapundalo, S.T., Hingorani, G.P., Haleen, S.J., LaDouceur, D.M., Hil, K.E., Flynn, M.A., and Reynolds, E.E. Structure–activity relationships of C-terminal endothelin hexapeptide antagonists. *J. Med. Chem.* 1992, **36,** 2585–2594

40 Doherty, A.M., Cody, W.L., He, J.X., DePue, P.L., Cheng, X.M., Welch, K.M., Flynn, M.A., Reynolds, E.E., LaDouceur, D.M., Davis, L.S., Keiser, J.A., and Hallen, S.J. In vitro and in vivo studies with a series of hexapeptide endothelin antagonists. *J. Cardiovasc. Pharmacol.* 1993, **22,** 98–102

41 Cody, W.L., Doherty, A.M., He, J.X., DePue, P.L., Rapundalo, S.T., Hingorani, G.P., Major, T.C., Panek, R.L., Dudley, D.T., Haleen, S.J., LaDouceur, D.M., Hill, K.E., Flynn, M.A., and Reynolds, E.E. Design of a functional hexapeptide antagonist of ET. *J. Med. Chem.* 1992, **35,** 3301–3303