Distributional Proteomics: Modelling Amino Acid Relationships by Measuring Their Patterns of Statistical Occurrence Across Proteins

Marcin Tatjewski[1,2] and Dariusz Plewczyński[2]

[1] Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
[2] Centre of New Technologies, University of Warsaw

Abstract.
Since 1990s the development of linguistic methods based on the *distributional hypothesis* has lead to significant achievements in extraction of word semantics from large corpora of texts in natural language. To date, there was no attempt made to apply algorithms of distributional semantics on biological data, despite many other successful method transfers between linguistic engineering and bioinformatics. Therefore, we constructed a distributional word-context matrix based on protein sequence data, making an analogy between words in natural language and single amino acids in proteins as most basic carriers of information. Our computational approach is inspired by the linguistic method of Correlated Occurrence Analogue to Lexical Semantics. In order to achieve our goal we also build a balanced set of protein sequences, as analogy to balanced text corpora in linguistics. The matrices which we obtained achieve correlations of up to 0.76 with amino acid substitution matrices. Substitution matrices are a widely used model of amino acid relationships, built using multiple sequence alignments and evolutionary data and useful in proteomics for sequence alignments. Our result suggests the potential to extract information about amino acids by purely statistical analysis of protein data. However, contrary to results in linguistic engineering, we obtain slightly higher correlation scores for matrices modelling simple tendency to co-occur than for matrices which model the more complex relationship of amino acids based on context.

1 Introduction

1.1 Linguistic distributional semantics

Linguistic distributional semantics is a part of a broader domain referred to as vector space models of semantics. This wide area aims at inferring word meaning from the statistical patterns of word usage in language. The foundations of the

field are build on theories like the *bag of words hypothesis*, the *latent relation hypothesis* or the *distributional hypothesis* [1]. The last one of them lies within the interest of our work. Distributional hypothesis states that the meaning of a word can be discovered by observation of the contexts in which the word occurs. Possibility of such inference is especially useful for detecting semantic similarity of words. It was already in 1950s that the *Distributional Hypothesis* was proposed [2]. However, it had to wait until the advent of computational methods in linguistic engineering before it could be applied on a larger scale.

The first proposed algorithm which used this hypothesis is Hyperspace Analogue to Language (HAL) [3]. It served its authors to construct some of the first word *semantic spaces*, also called *matrices of semantic relatedness* [4]. HAL method also established the four main steps which till now are the main ingredients of distributional semantics procedures. These steps are:

1. Gathering and preparing the text corpus which will serve as the experimental base. Linguists pay special attention to this stage and try to build *balanced* corpora. A balanced corpus includes texts from diverse sources: spoken language, books, newspapers, letters etc. An example of a balanced corpus is the balanced version of the National Corpus of Polish [5].
2. Processing the text corpus with a sliding window in order to obtain a co-occurrence count matrix. Sliding windows can vary in size, can be ramped or flat.
3. Post-processing of the obtained co-occurrence matrix, which at least should involve normalization, yet it may contain more advanced transformations or dimensionality reduction.
4. Establishing a similarity measure between the word-vectors described in the finally obtained matrix.

Soon after HAL appeared the better-known Latent Semantic Analysis (LSA) [6], yet it was a move from word-context model to word-document model. On the other hand, a continuation and extension of the HAL's word-context approach was proposed in the Correlated Occurrence Analogue to Lexical Semantics (COALS) model [7]. Concepts from the COALS algorithm were the most inspirational for our work. Supplement A presents an extract of our previous work: an example of word similarity results obtained with use of COALS method in a study of word synonymy for Polish language.

Nowadays, vector space models of semantics are a well developed domain with many successful applications. They are also easy to use, as there are many available text corpora and dedicated software packages [8].

Our experience of work with linguistic distributional semantics lead us in our daily bioinformatics research to an attempt to apply similar techniques in order to model relationships (*semantics*) between amino acids.

1.2 Amino acid relationship modelling

Modelling differences between amino acids is an important task for proteomics. Whether we want to extract features for machine learning from amino acid

sequences or whether we attempt to align several proteins in order to reason about their common traits, we need to be able to quantitatively compare amino acids to each other. Several resources are available for such tasks. Many of them are gathered in the AAIndex Database [9]. Two types of such resources which we would like to cover are:

Amino acid indices - These are mostly physicochemical properties with specific values for each amino acid. They can relate to hydrophobicity, compositional properties, structural propensity, electric properties and others. Indices are useful, for example, in the task of feature extraction from protein sequences [10]. Their abundance might pose a challenge, yet this is often addressed with clustering or feature selection methods [11].

$$amino\ acid\ index : AA \rightarrow R$$

Equation 1: Amino acid indices - functions returning values of physicochemical attributes for each amino acid. Below AA is the set of amino acids, and R is the set of real values.

Substitution (or mutation) matrices - The idea behind them comes from the need to align protein sequences with each other. Before their appearance, alignment scoring algorithms counted sequence matches or mismatches equally - not taking into account which particular amino acids are compared. Later on, based on the observation that some amino acids are more likely to mutate than others (and also mutate to specific targets), biologists started to differently rate the proximity of protein sequences. To establish the substitution scores, scientist relied on analysis of multiple alignments. In PAM aligned were evolutionary similar proteins [12], while in BLOSUM alignment focused on very conserved regions in distant proteins [13]. Other methods were also designed, yet these two substitution matrices are one of the most popular and are commonly used in the popular BLAST program [14].

$$substitution\ matrix : AA \times AA \rightarrow R$$

Equation 2: Substitution matrices - functions returning evolutionary/chemical similarity/dissimilaity scores for pairs of amino acids. Below AA is the set of amino acids, and R is the set of real values.

Substitution matrices in proteomics serve similar purpose as semantic spaces in linguistic. They rate proximity of, respectively, amino acids and words. However, both methods are based on a very different approach. In our work we decided to apply the methods of distributional semantics to proteins, thus building

an *amino acid semantic space*. We base this approach on the analogy between words and amino acids - both are the basic units carrying information in their domains and both occur in large sequences that can be studied statistically.

2 Method and materials

2.1 Protein corpus preparation

As we mentioned in section 1.1, an important step of every linguistic engineering experiment is careful preparation of a text corpus. Therefore, we decided to pay special attention to this step in our biological experiment. Taking just a full set of proteins from a database like UniProt might have biased the results, as proteins in different domains are not equally well researched. For some types of organisms or some functional types of proteins we have much more objects sequenced than in other areas.

In order to obtain a balanced protein set we utilized UniProt20 database which was developed in the HHblits package [15]. UniProt20 is a clustered set of proteins from UniProt. It was constructed using similarity threshold of 20%. To build our protein corpus we took at maximum one sequence from each cluster of UniProt20. However, we only accepted sequences that are marked in the original UniProt as having experimental evidence at protein or transcript level [16]. Therefore, some UniProt20 clusters are not at all represented in our dataset. Statistics of the protein corpus which we obtained are displayed in Table 2.1.

Number of sequences	347 409
Number of amino acids	101 966 845
Mean sequence length	293.5
Median sequence length	193.0

Table 1. Statistics of the obtained protein corpus.

In order to make sure that short, medium and long sequences are relatively equally represented in our dataset we looked in detail into its composition from the perspective of elements' length, what is displayed in Figures 1 and 2. Moreover, Figure 3 presents the amino acid composition of our dataset.

2.2 Amino acid distributional matrix construction

Procedure which we used to construct our amino acid distributional matrix is highly inspired by the COALS algorithm [7]. However, many of steps in COALS are appropriate only for linguistic domain, thus we eliminated:

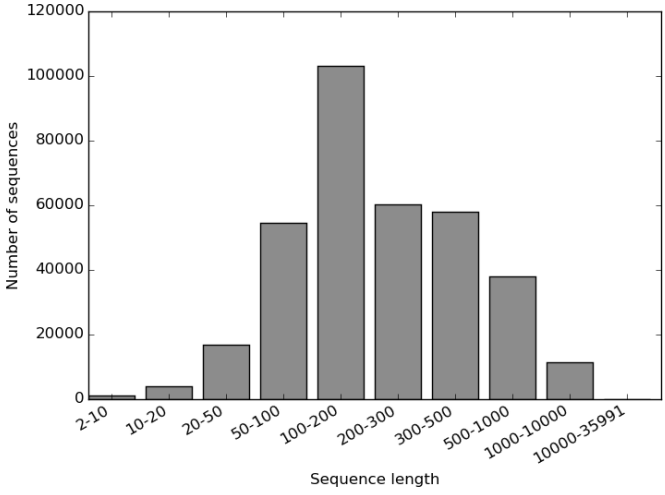

Fig. 1. Number of sequences in the protein corpus per sequence length group

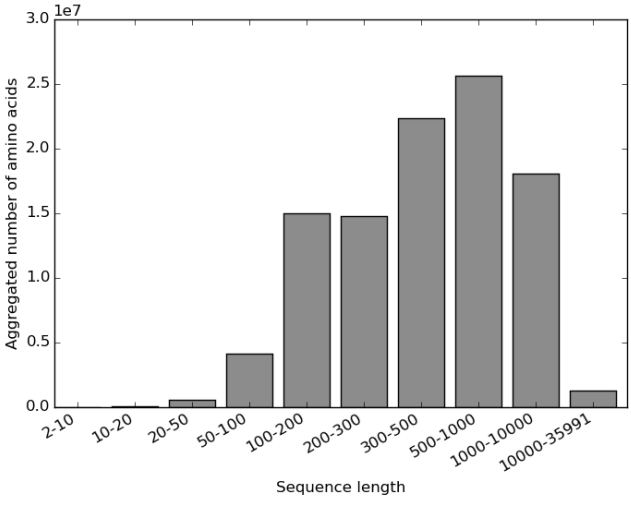

Fig. 2. Aggregated number of amino acids in the protein corpus per sequence length group

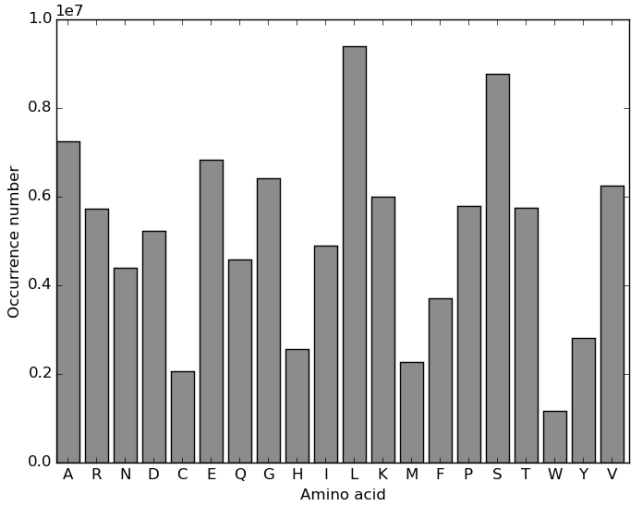

Fig. 3. Distribution of different amino acids in the protein corpus

– Dimensionality reduction, which is required in linguistics because of the huge size of the matrix based on words and is also credited with performance increase. Reduction of our 20×20 amino acid matrix does not seem to be necessary, yet it might be worth checking in the future whether it would not increase the final performance.
– Replacing negative correlation values with zeros. In linguistics this transformation step is explained by the fact that knowing a large set of words that have negative correlation with a target word is not very helpful for inferring the meaning of the target word. On the other hand, knowledge of just a handful of words that correlate positively with the target word provides a lot of insight into the target word's semantics. For example, information that an unknown word W has negative correlation with words: mountain, swimming, colorful, multiple and clumsy is not very useful when we want to infer the semantics of W. However, if we know that W correlates positively with words: dog, lion and pet, than we know much more about its meaning. Nevertheless, the world of amino acids is different. We cannot claim *a priori* that negative correlation between amino acids is meaningless. Intuitively it's seems to be quite the opposite. Therefore, we keep negative correlation values as equally valuable as positive correlations.

Therefore, our amino acid procedure consisted of the following steps:

1. Gathering co-occurrence counts in matrix of size 20×20 using a sliding window. We used flat and ramped windows with radius 4,10 and 16, thus obtaining 6 different matrices. Figure 2.2 shows how a ramped window of

radius 4 counts co-occurrence scores. The size of our final matrix is 20×20 as we decided to ignore all the non-standard amino acids as their number was not big enough.
2. Co-occurrence matrix normalization with use of formula from Figure 3. The formula should not be confused with correlation of two co-occurrence rows, as it is instead a correlation between the occurrences of two amino acids [7].
3. Obtaining similarity score for two amino acids by calculating correlation between their row-vectors. This step shifts the final results from looking at pure co-occurrence likelihood of amino acids towards the representation of their context similarity.

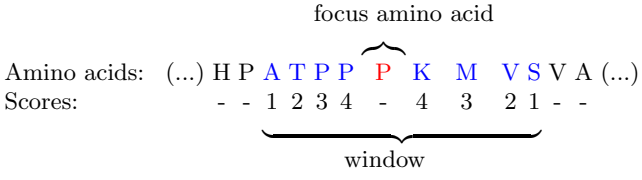

Fig. 4. Co-occurrence scoring for a ramped window with radius 4 on an example protein subsequence.

$$D_{i,j} = \frac{S * C_{a,b} - \sum_{j}^{20} C_{a,j} * \sum_{i}^{20} C_{i,b}}{(\sum_{j}^{20} C_{a,j} * (S - \sum_{j}^{20} C_{a,j}) * \sum_{i}^{20} C_{i,b} * (S - \sum_{i}^{20} C_{i,b}))^{1/2}}$$

where

$$S = \sum_{i}^{20} \sum_{j}^{20} C_{i,j}$$

Equation 3: Transforming raw co-occurrence counts (C matrix) into distribution-based tendency to co-occur (D matrix) through normalization based on Pearson's correlation coefficient. 20 is the number of basic amino acids, thus it determines the dimensionality of matrices C and D.

2.3 Comparison with substitution matrices

The most intuitive idea for checking whether the information gathered in our matrices does make biological sense is to compare them to substitution matrices.

Following works on comparing biological matrix resources with each other, we decided to use for this task a simple correlation of matrices flattened to rows, as presented in Figure 4 [17]. For comparison we took all the 93 matrices available in AAIndex2 Database [9]. It's important to note, that substitution matrices gathered in this resource are not all very different from each other, as this set contains many variations of matrices built by the same algorithms. An example of a substitution matrix is presented in Table 2.3.

For comparison with substitution matrices we did not only take our final row similarity matrices, but we also performed calculations for matrices which we obtain before performing step number 3 from procedure presented in section 2.2.

A	5																		
R	-2	7																	
N	-1	0	6																
D	-2	-1	2	7															
C	-1	-3	-2	-3	12														
E	-1	0	0	2	-3	6													
Q	-1	1	0	0	-3	2	6												
G	0	-2	0	-1	-3	-2	-2	7											
H	-2	0	1	0	-3	0	1	-2	10										
I	-1	-3	-2	-4	-3	-3	-2	-4	-3	5									
L	-1	-2	-3	-3	-2	-2	-2	-3	-2	2	5								
K	-1	3	0	0	-3	1	1	-2	-1	-3	-3	5							
M	-1	-1	-2	-3	-2	-2	0	-2	0	2	2	-1	6						
F	-2	-2	-2	-4	-2	-3	-4	-3	-2	0	1	-3	0	8					
P	-1	-2	-2	-1	-4	0	-1	-2	-2	-2	-3	-1	-2	-3	9				
S	1	-1	1	0	-1	0	0	0	-1	-2	-3	-1	-2	-2	-1	4			
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-1	-1	2	5		
W	-2	-2	-4	-4	-5	-3	-2	-2	-3	-2	-2	-2	-2	1	-3	-4	-3	15	
Y	-2	-1	-2	-2	-3	-2	-1	-3	2	0	0	-1	0	3	-3	-2	-1	3	8
V	0	-2	-3	-3	-1	-3	-3	-3	-3	3	1	-2	1	0	-3	-1	0	-3	-1	5
	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V

Table 2. Example of a substitution matrix: BLOSUM45 substitution matrix (Henikoff-Henikoff, 1992). Missing values above the diagonal indicate that the matrix is symmetric. Please note that not all substitution matrices are symmetric.

3 Results and discussion

Our amino acid distributional matrices obtain surprisingly high correlations with the substitution matrices, e.g. 0.76 with matrix built by *Koshi et al*, 0.64 with BLOSUM45 or 0.51 with PAM120. It's especially interesting as our matrices are built using a very different paradigm. Distributional amino acid matrices

$$c = \frac{\sum_{i}^{N}\sum_{j}^{N}(D_{i,j} - \bar{D})(S_{i,j} - \bar{S})}{(\sum_{i}^{N}\sum_{j}^{N}(D_{i,j} - \bar{D})^2 \sum_{i}^{N}\sum_{j}^{N}(S_{i,j} - \bar{S})^2)^{1/2}}$$

Equation 4: Matrix correlation by flattening matrices to vectors. D - distributional amino acid matrix. S - substitution matrix.

are based on vertical analysis of protein sequences and they do not use any external knowledge about evolutionary relationships between proteins. On the other hand, most of the substitution matrices rely on horizontal analysis of protein multiple alignments and incorporate evolutionary information into their methodology. Also notable is the result of 0.73 correlation with a substitution matrix based on amino acid chemical properties [18]. These results show that it is possible to extract meaningful knowledge about amino acids from pure statistical analysis of protein sequences.

However, it's important to note that, contrary to the results in linguistic applications, better "performance" is achieved by distributional matrix built without the step 3 presented in method from section 2.2. This means that more related to substitution matrices is the pure likelihood of amino acid co-occurrence rather than semantic similarity driven by the context relationship.

Acknowledgements

The study is cofounded by the European Union from resources of the European Social Fund. Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00; Polish National Science Centre (grant numbers: 2015/16/T/ST6/00493, 2014/15/B/ST6/05082 and 2013/09/B/NZ2/00121); EU COST BM1405 and BM1408 actions.

A Supplement: Example distributional semantics results for Polish language

Tables in Figure A present an example of distributional semantics results for Polish language. Usually, most valued and useful outcomes of these methods are lists of words' nearest neighbors, i.e. words having most similar vectors in the semantic space to a given word. In the case of Figure A we see nearest neighbors lists from space produced with COALS algorithm [7] run on the National Corpus of Polish [5]. Results were produced for the project APPROVAL[1], which was aimed at analysis of synonym pairs. This is why we present neighbors lists for two synonymous words [28].

[1] http://www.approval.uw.edu.pl/start

Substitution matrix AAIndex code	Correlation with matrix built without step 3. from section 2.2	Correlation with matrix built including step 3. from section 2.2	Substitution matrix description
KOSJ950102	**0.76**	0.58	Context-dependent optimal substitution matrices for exposed beta [19]
OVEJ920105	0.75	0.58	Environment-specific amino acid substitution matrix for inaccessible residues [20]
LINK010101	0.75	0.58	Substitution matrices from a neural network model [21]
MCLA720101	0.73	**0.68**	Chemical similarity scores [18]
CSEM940101	0.67	0.65	Residue replace ability matrix [22]
HENS920101	0.64	0.58	BLOSUM45 substitution matrix [13]
ALTS910101	0.51	0.47	The PAM-120 matrix [23]
AZAE970102	0.45	0.4	The substitution matrix derived from spatially conserved motifs [24]
GEOD900101	0.33	0.33	Hydrophobicity scoring matrix [25]
RUSR970101	-0.01	0.04	Substitution matrix based on structural alignments of analogous proteins [26]
GRAR740104	-0.43	- 0.41	Chemical distance [27]

Table 3. Correlation results (calculated according to Figure 4) of chosen substitution matrices with amino acid distributional matrices based on co-occurrences calculated by flat sliding window of radius 16.

Kolarz	
Similarity	**Neighbor**
0.859	biegacz
0.783	kolarski
0.775	maratończyk
0.766	kajakarz
0.762	peleton
0.761	lekkoatleta
0.758	pływak
0.755	wyścig
0.744	zawodnik
0.736	rajdowiec
0.727	szosowy
0.726	szachista
0.719	zapaśnik
0.717	kolarstwo
0.716	jaskuła
0.707	olimpijczyk

Cyklista	
Similarity	**Neighbor**
0.767	motocyklista
0.761	rowerzysta
0.715	pieszy
0.705	rajdowiec
0.704	rowerowy
0.688	kolarz
0.683	jednoślad
0.668	rower
0.667	zmotoryzowany
0.661	rajd
0.647	motocyklowy
0.641	biegacz
0.638	quad
0.624	kolarski
0.624	motocykl
0.621	spacerowicz

Fig. 5. Example nearest neighbors lists from a COALS [7] semantic space constructed over National Corpus of Polish [5] in the course of project APPROVAL (http://www.approval.uw.edu.pl/start).

References

1. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. Journal of Artificial Intelligence Research **37**(1) (January 2010) 141–188
2. Harris, Z.: Distributional structure. Word **10**(23) (1954) 146–162
3. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers **28**(2) (June 1996) 203–208
4. Piasecki, M., Szpakowicz, S., Broda, B.: A Wordnet from the Ground Up. (2009)
5. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B.: Narodowy Korpus Języka Polskiego. (2012)
6. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes **25**(2-3) (January 1998) 259–284
7. Rohde, D., Gonnerman, L., Plaut, D.: An improved model of semantic similarity based on lexical co-occurrence. Communications of the ACM **8** (2006) 627–633
8. Jurgens, D., Stevens, K.: The S-Space Package: An Open Source Package for Word Space Models. Proceedings of the ACL 2010 System Demonstrations (July) (2010) 30–35
9. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.: AAindex: Amino acid index database, progress report 2008. Nucleic Acids Research **36**(SUPPL. 1) (2008) 202–205
10. Plewczynski, D., Basu, S., Saha, I.: AMS 4.0: consensus prediction of post-translational modifications in protein sequences. Amino acids **43**(2) (August 2012) 573–82
11. Saha, I., Maulik, U., Bandyopadhyay, S., Plewczynski, D.: Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino acids **43**(2) (August 2012) 583–94
12. Dayhoff, M., Schwartz, R.: A Model of Evolutionary Change in Proteins. In Atlas of protein sequence and structure (1978) 345–352
13. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America **89**(22) (November 1992) 10915–9
14. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3) (October 1990) 403–10
15. Remmert, M., Biegert, A., Hauser, A., Söding, J.: HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods **9**(2) (February 2012) 173–5
16. The UniProt Consortium: UniProt: a hub for protein information. Nucleic Acids Research **43**(D1) (October 2014) D204–D212
17. Tomii, K., Kanehisa, M.: Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein engineering **9**(1) (1996) 27–36
18. McLachlan, A.D.: Repeating sequences and gene duplication in proteins. Journal of molecular biology **64**(2) (March 1972) 417–37
19. Koshi, J.M., Goldstein, R.A.: Context-dependent optimal substitution matrices. Protein engineering **8**(7) (July 1995) 641–5
20. Overington, J., Donnelly, D., Johnson, M.S., Sali, A., Blundell, T.L.: Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein science : a publication of the Protein Society **1**(2) (February 1992) 216–26

21. Lin, K., May, A.C., Taylor, W.R.: Amino Acid Substitution Matrices from an Artificial Neural Network Model. Journal of Computational Biology **8**(5) (October 2001) 471–481
22. Cserzö, M., Bernassau, J.M., Simon, I., Maigret, B.: New alignment strategy for transmembrane proteins. Journal of molecular biology **243**(3) (October 1994) 388–96
23. Altschul, S.F.: Amino acid substitution matrices from an information theoretic perspective. Journal of molecular biology **219**(3) (June 1991) 555–65
24. Azarya-Sprinzak, E., Naor, D., Wolfson, H.J., Nussinov, R.: Interchanges of spatially neighbouring residues in structurally conserved environments. Protein engineering **10**(10) (October 1997) 1109–22
25. George, D.G., Barker, W.C., Hunt, L.T.: Mutation data matrix and its uses. Methods in enzymology **183** (January 1990) 333–51
26. Russell, R.B., Saqi, M.A., Sayle, R.A., Bates, P.A., Sternberg, M.J.: Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. Journal of molecular biology **269**(3) (June 1997) 423–39
27. Grantham, R.: Amino acid difference formula to help explain protein evolution. Science (New York, N.Y.) **185**(4154) (September 1974) 862–4
28. Tatjewski, M., Bańko, M., Kucińska, A., Rączaszek-Leonardi, J.: Computational distributional semantics and free associations: a comparison of two word-similarity models in a study of synonyms and lexical variants. In Pęzik, P., Waliński, J., Kosecki, K., eds.: Language, Corpora and Cognition. Peter Lang (In press)