Optimizing Property Codes in Protein Data Reveals Structural Characteristics

Olaf Weiss[1⋆], Andreas Ziehe[1], and Hanspeter Herzel[2]

[1] Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany
[2] Institute for Theoretical Biology, Humboldt-University Berlin Invalidenstr. 43, 10115 Berlin, Germany

Abstract. We search for assignments of numbers to the amino acids (property codes) that maximize the autocorrelation function signal in given protein sequence data by an iterative method. Our method yields similar results to optimization with the related extended Jacobi method for joint diagonalization and standard optimization tools.
In nonhomologous sets representative of all proteins we find optimal property codes that are similar to hydrophobicity but yield much clearer correlations. Another property code related to α-helix propensity plays a less prominent role representing a local optimum. We also apply our method to sets of proteins known to have a high content of α- or β-structures and find property codes reflecting the specific correlations in these structures.

Introduction

Property codes such as hydrophobicities are used in a wide range of bioinformatical applications such as prediction of transmembrane regions [1], secondary structure prediction [2], or derivation of contact potentials for protein 3D structure modeling [3]. Furthermore, property codes are necessary to translate protein sequences into time series which can be analyzed by artificial neural networks [4, 5], correlation functions [6] and other methods.

The AAindex database [7] currently contains 437 property codes, many of which are related. The choice of which property code to use is rather arbitrary.

A few studies have been aimed at optimizing property codes from mutation matrices [8, 9] or 3D-structure modelling [3, 5].

In this study we aim at finding property codes leading to a large autocorrelation function signal strength in given protein sequence data. We have already addressed this question in [10] by brute force random search. In this study, we present a fast iterative method based on matrix diagonalization. Furthermore, the similar concept of joint Matrix diagonalization with Jacobi methods [11] is also applied. To our knowledge, this is the first use of this method in a bioinformatics context. We compare our results to off-the-shelf optimizers and get

⋆ corresponding author, e-mail: olaf.weiss@first.fraunhofer.de

similar results. However, our method has the advantage of discovering additional biologically relevant property codes from local optima.

Finally we show how our method can be applied to sets of protein sequences rich in specific secondary structure. We find property codes and autocorrelation patterns specific to α-helices and β-strands.

Methods

We consider property codes as a translation of the 20 amino acids to numbers. We write a property code as a vector a, whose elements are the values that the code assigns to the amino acids. A property code represents one possible mapping of protein sequences to numerical series upon which time series analysis tools can be applied. We denote such a resulting numerical series by $(x_l^{(\mathsf{a})})$.

The Autocorrelation Function in Symbol Sequences

The autocorrelation function (acf) of a protein sequence using the property code a

$$C_{\mathsf{a}}(k) =< x_l^{(\mathsf{a})} x_{l+k}^{(\mathsf{a})} > - < x_l^{(\mathsf{a})} >< x_{l+k}^{(\mathsf{a})} > \tag{1}$$

can also be written as a quadratic form

$$C_{\mathsf{a}}(k) = \mathsf{a}^t \mathbf{D}(k) \mathsf{a} \tag{2}$$

of the matrix $\mathbf{D}(k)$ whose elements are defined by

$$D_{ij}(k) = P_{ij}(k) - p_i q_j \tag{3}$$

where $P_{ij}(k)$ is the joint probability of finding residues i and j separated by $k-1$ positions, $p_i = \sum_j P_{ij}(k)$, and $q_j = \sum_i P_{ij}(k)$ [12]. Thus, the contribution of the sequence to the acf is captured in $\mathbf{D}(k)$. Note that the natural estimator of the $D_{ij}(k)$ is biased and needs to be corrected [10]. A plot of the Kyte-Doolittle hydrophobicity [13] acf using (2) averaged over the `pdb_select set` of protein sequences [14] is plotted as the black line in Fig. 1.

As already done in [10], we define the signal strength

$$S_{\mathsf{a}} := \sum_{k=1}^{k_{\max}} \left(C_{\mathsf{a}}(k)\right)^2 \tag{4}$$

as a quantity that we want to maximize by varying a to find optimal property codes. As constraint we fix

$$\mathsf{a}^t \mathsf{a} = \sum_i a_i^2 = 1. \tag{5}$$

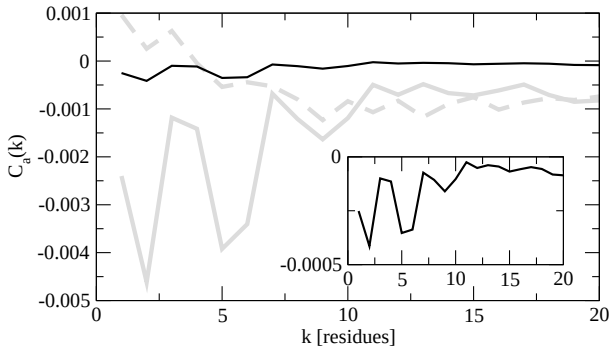

Fig. 1. Acfs in the `pdb_select` data: Kyte-Doolittle hydrophobicity-acf (*black line, also enlarged in the insert*) and optimized acfs with signal strength $S = 68 \cdot 10^{-6}$ (*grey solid*) and $S = 13 \cdot 10^{-6}$ (*grey dashed*)

The Eigenvector Iteration

Maximizing $C_{\mathsf{a}}(k)$ for one single k with constraint (5) leads to the eigenvalue equation

$$\mu \mathsf{a} = \mathbf{D}^{sym}(k)\mathsf{a} \tag{6}$$

where $\mathbf{D}^{sym} = \frac{1}{2}(\mathbf{D} + \mathbf{D}^t)$. The eigenvalues $\mu^i = C_{\mathsf{a}^i}(k)$ are the acf of the eigenvectors a^i. This can also be applied to maximize a linear combination $\sum_k \xi_k C_{\mathsf{a}}(k) = \mathsf{a}^t \mathbf{\Delta} \mathsf{a}$ with

$$\mathbf{\Delta} = \sum_{k=1}^{k_{\max}} \xi_k \mathbf{D}^{sym}(k), \tag{7}$$

because C_{a} is linear in the $\mathbf{D}(k)$.

From this we derived the following iteration which puts emphasis on those k where the acf $C_{\mathsf{a}}(k)$ is far away from zero:

1. Estimate the corrected covariance matrices $\mathbf{D}(k)$ from sequence data.
2. Set ξ_k as an initialization.
3. Calculate the matrix $\mathbf{\Delta}$ as in (7) using the current ξ_k.
4. Maximize $\mathsf{a}^t \mathbf{\Delta} \mathsf{a}$ by finding the eigenvector corresponding to the largest eigenvalue of the symmetric part of $\mathbf{\Delta}$.
5. Change the ξ_k to $\xi_k = C_{\mathsf{aa}}(k)$ with the property code a obtained in step 4.
6. Return to step 3 with the new ξ_k.

A mathematical analysis of this iteration procedure shows that it *i)* finds a local maximum of S, and *ii)* is closely related to gradient methods for optimization (see [15, appendix A]). We termed this iteration *eigenvector iteration*. It usually converges after 15–25 steps.

To ensure that we find all relevant local maxima, we perform the iteration with 1000 random initial conditions. This takes a few minutes on a PC. In Fig. 1

the effect of the optimization can be seen: the solid grey line shows the acf of an optimized hydrophobicity which has a much larger signal of similar structure to that of the Kyte-Doolittle hydrophobicity acf (black line and insert).

Extended Jacobi method for joint diagonalization

An alternative method to determine an eigenvector (property code) corresponding to the maximal (average) eigenvalue is based on joint diagonalization of the correlation matrices $\mathbf{D}(k)$ as defined in eq. (3). Instead of forming the weighted sum of $\mathbf{D}(k)$ as in (7) and solving the eigenvalue equation we propose using a simultaneous diagonalization method by Cardoso and Souloumiac (1996)[11]. Exact joint diagonalization is formally defined for a set of normal matrices $\mathbf{M}_1, \ldots, \mathbf{M}_n$ which commute and means that an orthogonal matrix $\mathbf{V}$ exists, such that $\mathbf{V}^T\mathbf{M}_1\mathbf{V}, \ldots, \mathbf{V}^T\mathbf{M}_n\mathbf{V}$ are all diagonal.

In practical applications an exact diagonalization is not possible, but one can still try to minimize the deviation from diagonality. Using the common diagonality measure

$$\mathrm{off}(\mathbf{M}) := \sum_{i \neq j} (M_{ij})^2,$$

leads to the following constrained optimization problem:

$$\min_{\mathbf{V}:\mathbf{V}^T\mathbf{V}=\mathbf{I}} \sum_{k=1}^{k_{max}} \mathrm{off}(\mathbf{V}^T\,\mathbf{M}_k\,\mathbf{V}).$$

The basic idea of the extended Jacobi technique is to approximate the orthogonal matrix $\mathbf{V}$ by applying a sequence of elementary rotations $\mathbf{R}_n(\phi_n)$ which minimize the off-diagonal elements at position (i, j) of the respective $\mathbf{M}_k$ matrices. One obtains the final solution by forming the product of all elementary rotations, i.e. $\mathbf{V} = \prod_n \mathbf{R}_n(\phi_n)$. It has been shown that the optimal rotation angle ϕ_n can be calculated in closed form, which leads to a highly efficient numerical algorithm[3] (for details see [11]).

For our application, we jointly diagonalize symmetrized $\mathbf{D}(k)$ matrices $\mathbf{D}^{sym}(k)$. Then we choose the eigenvector with the highest average eigenvalues of the $\mathbf{D}^{sym}(k)$.

While this method has been applied successfully to real world problems such as blind source separation [16, 17], this represents – to the best of our knowledge – the first application of joint diagonalization in bioinformatics.

Optimization with standard tools

We also numerically optimized $S_{\mathtt{a}}$ as a function of the a_i using the MATLAB optimization toolbox. The off-the-shelf method applied – `fmincon()` with medium-scale optimization – is based on Sequential Quadratic Programming and line search as described in [18]. The results obtained confirm the maxima found with the other algorithms described above.

[3] see `http://tsi.enst.fr/~cardoso/jointdiag.html` for MATLAB code

Results

We applied the different methods described above to several sets of protein sequences. The resulting optimized property codes are listed in Tab. 1.

Table 1. Property Codes: The amino acids are ordered according to the optimized hydrophobicity in the `pdb_select set`. S_a $[10^{-6}]$ gives the signal strength of the property code in the data it was optimized with.

	p_s 1	p_s 2	jdiag	matl.	α 1	α 2	β 1	β 2	β 3
L	-0.63	-0.19	0.68	-0.64	-0.16	-0.62	-0.41	0.20	0.25
I	-0.34	0.00	0.34	-0.35	0.14	-0.26	-0.21	0.09	-0.08
V	-0.28	0.04	0.18	-0.27	-0.13	-0.31	-0.48	0.13	-0.15
F	-0.19	0.13	0.20	-0.20	0.03	-0.21	-0.19	0.16	-0.02
Y	-0.09	0.07	0.09	-0.09	0.09	-0.11	-0.08	0.06	-0.25
W	-0.07	-0.01	0.07	-0.07	-0.01	-0.03	-0.05	0.06	-0.04
M	-0.06	-0.11	0.06	-0.06	0.00	-0.09	-0.06	0.06	-0.02
C	-0.02	0.04	0.01	-0.01	0.09	-0.09	-0.01	0.03	-0.03
A	-0.01	-0.56	-0.01	0.00	-0.82	0.12	0.14	0.24	0.82
H	0.03	0.21	-0.04	0.03	0.04	-0.04	0.01	-0.12	-0.08
T	0.04	0.09	-0.03	0.04	0.06	0.09	-0.16	-0.26	-0.07
P	0.08	0.32	-0.07	0.08	0.08	0.05	0.11	0.02	-0.03
R	0.10	-0.10	-0.15	0.11	0.08	0.16	0.02	-0.03	0.06
Q	0.12	-0.13	-0.15	0.13	-0.10	0.18	-0.01	-0.23	-0.08
S	0.13	0.44	-0.11	0.12	0.17	0.10	0.16	-0.45	0.14
N	0.17	0.03	-0.15	0.17	0.08	0.12	0.19	-0.16	-0.25
K	0.19	-0.23	-0.21	0.20	-0.22	0.31	0.00	0.06	0.08
G	0.20	0.25	-0.11	0.18	0.31	0.06	0.46	0.61	-0.21
D	0.30	0.07	-0.25	0.29	0.20	0.24	0.39	-0.22	-0.11
E	0.32	-0.36	-0.35	0.32	0.07	0.33	0.18	-0.23	0.08
S_a $[10^{-6}]$	68	13	67	68	49	74	49	30	38

Application to the `pdb_select` data

As a representative set of proteins we took the current version (Apr. 2002) of the `pdb_select` list [14], which comprises 1771 proteins of less than 25% sequence identity from the Brookhaven Protein Database. We computed $\mathbf{D}(k)$ in the 1414 sequences longer than 70 residues after randomly assigning one of the 20 amino acids to the undetermined positions marked by X in the sequence.

The eigenvector iteration gives an optimized correlation function (grey, solid line in Fig. 1) with a similar shape as the hydrophobicity acf, but with an amplitude that is an order of magnitude larger. The corresponding property code (Tab. 1, col. 'p_s 1') has some similarity to hydrophobicities, the correlation coefficient ρ with the Kyte-Doolittle hydrophobicity is 0.46, that with the Goldmann-Engelmann-Steitz (GES) hydrophobicity [19] is 0.66. This property code can be seen as an optimized hydrophobicity. The oscillation in the acf corresponds to amphiphilic α-helices with the hydrophobic side facing to the core of the protein and the polar side exposed to the solvent [6]. This optimum is also found using the joint diagonalization method (Tab. 1, col. 'jdiag') and the MATLAB optimization toolbox (col. 'matl.').

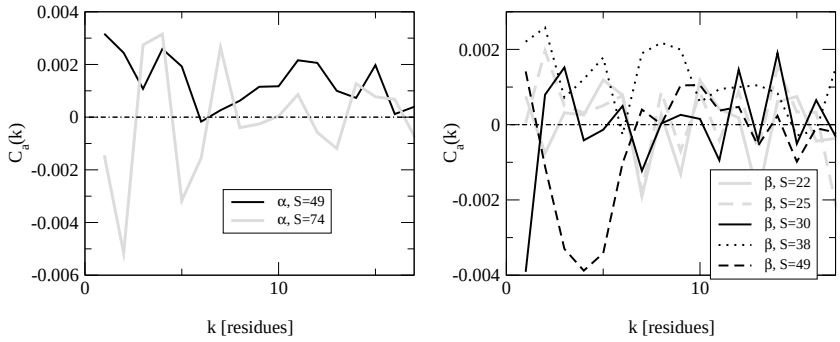

Fig. 2. The acfs using property codes optimized in the sets of α-helix (*left*) and β-strand rich proteins (*right*) in the respective sets

With some initializations the eigenvector optimization finds another local maximum of S_{a} which is related to α-helix propensity (Tab. 1, col. 'p_s 2'). The corresponding acf (dashed grey line, Fig. 1) shows a nonoscillating decay with the same length scale as α-helices, as described in [10]. This local optimum is reached from only about 0.05% of the initial conditions. It is one of the advantages of the eigenvector iteration that it is capable of finding other local optima than the highest one. As shown for this data, additional local maxima are also of biological relevance as α-helix-propensity is a helpful property for secondary structure prediction [20].

Secondary Structure

We also performed the eigenvector iteration in sets of sequences of proteins having mainly one type of secondary structure. We use the α-helix- and β-strand-rich datasets introduced in [10] comprising less than 100 sequences each. Therefore, all correlation functions estimated in these sets have a high level of noise. Still, correlation signals can be measured in these sets. However, the optimization procedures are facing rather rough landscapes where multiple local maxima are to be expected.

There are two local optima in the set of sequences from α-helix rich proteins. The corresponding acfs are plotted in the left graph of Figure 2. The more weakly correlated property causes a signal strength of $S = 49 \cdot 10^{-6}$ in the autocorrelation function (black line, col. α 1 in Tab. 1). Although the acf is heavily affected by fluctuations it shows a clear tendency towards positive correlations, especially for $k < 5$. Indeed, this property is weakly related to α-helix propensity ($\rho = -0.55$) while it is uncorrelated to the GES hydrophobicity scale ($\rho = 0.10$). Thus this property constitutes an optimized α-helix propensity. The other optimized property code in the α-helical protein sequence set is clearly correlated to hydrophobicity ($\rho = 0.71$). It is therefore not surprising to see that the corresponding acf strongly oscillates. It could be termed α-helix specific hydrophobicity as it is optimized to detect the oscillations caused by α-helices.

The picture in the set of β-strand rich proteins is more confusing as there are as many as five local optima. The two optimized property codes that cause the weakest signal ($S = 22 \cdot 10^{-6}$ and $S = 24 \cdot 10^{-6}$) are only reached by very few initial conditions. The corresponding acfs look very noisy (grey lines in the right graph of Fig. 2). We consider that these property codes are results of optimizing the noise rather than true (i.e. biological) signals.

The three other optimized property codes show weak correlations to hydrophobicity (those with $S = 49 \cdot 10^{-6}$ and $30 \cdot 10^{-6}$) or α-helix propensity (the one with $S = 38 \cdot 10^{-6}$, cols. β 1–3 in Tab. 1). There are two prominent features visible in the corresponding acfs. The property with $S = 30 \cdot 10^{-6}$ shows a clear anticorrelation at $k = 1$. This means that the corresponding property has the tendency to be concentrated on one side of the β-sheet. An even clearer signal can be seen in the acf of the property with $S = 49 \cdot 10^{-6}$: a pronounced negative peak at $k = 3, 4, 5$. The exact biochemical explanation for this is still unclear, but the distance is in the typical range of short loops connecting antiparallel β-strands. Thus, this signal could be caused by residues that actually have physical contact in the protein. Presumably certain residue combinations are selected to stabilize short loops (i.e. turns) in β-sheets.

Discussion

For given sets of protein sequences, we have maximized the correlation function signal by varying the property code, i.e. translation of the sequences to numerical series. We have done this using our eigenvector iteration, extended Jacobi method for joint diagonalization – introduced here for the first time in bioinformatics – and an off-the-shelf MATLAB optimization. All methods yielded similar results, the eigenvector iteration can find additional local optima which often hold additional biological significance.

Hydrophobicity- and α-helix propensity-related properties are found in the `pdb_select` dataset representing all proteins. These types of properties are found to be more pronounced in a set of α-helix-rich proteins. The optimized properties in β-strand-rich proteins include a novel correlation pattern with anticorrelations at $k = 3 \ldots 5$.

Our methods are applicable to other sets of proteins as well, e.g. comparisons of optimized property codes in proteomes of different organisms will be done in forthcoming studies.

Acknowledgements

We thank the EU for funding via project IST-1999-14190 – BLISS and the DFG for funding via SFB 618. Furthermore, we gratefully acknowledge Johannes Schuchhardt, Sebastian Mika, and David Tax for help- and fruitful discussions, as well as Caspar van Wrede and Klaus-Robert Müller for proofreading.

References

1. Eisenberg, D., Schwarz, E., Komaromy, M., Wall, R.: Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. **179** (1984) 125–142
2. Frishman, D., Argos, P.: Seventy-five percent accuracy in protein secondary structure prediction. Proteins **27** (1997) 329–335
3. Casari, G., Sippl, M.J.: Structure-derived hydrophobic potential. J. Mol. Biol. **224** (1991) 725–732
4. Jagla, B., Schuchhardt, J.: Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. Bioinformatics **16** (2000) 245–250
5. Lin, K., May, A.C.W., Taylor, W.: Amino acid encoding schemes from protein structure alignments: Multi-dimensional vectors to describe residue types. J. theor. Biol. **216** (2002) 361–365
6. Kanehisa, M.I., Tsong, T.Y.: Hydrophobicity and protein structure. Biopolymers **19** (1980) 1617–1628
7. Kawashima, S., Kanehisa, M.: AAindex: Amino acid index database. Nucleic Acids Res. **28** (2000) 374
8. Sneath, P.H.A.: Relations between chemical and biological activity in peptides. J. theor. Biol **12** (1966) 157–195
9. Stanfel, L.E.: A new approach to clustering the amio acids. J. theor. Biol. **183** (1996) 195–205
10. Weiss, O., Herzel, H.: Correlations in protein sequences and property codes. J. theor. Biol. **190** (1998) 341–353
11. Cardoso, J.F., Souloumiac, A.: Jacobi angles for simultaneous diagonalization. SIAM J. Mat. Anal. Appl. **17** (1996) 161–164
12. Herzel, H., Große, I.: Measuring correlations in symbol sequences. Physica A **216** (1995) 518–542
13. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157** (1982) 105–132
14. Hobohm, U., Sander, C.: Enlarged representative set of protein structures. Protein Sci. **3** (1994) 552–554
15. Weiss, O.: Korrelationen und Eigenschaftscodes in Proteinsequenzen (Correlations and Property Codes in Protein Sequences). Logos, Berlin (2001) PhD Thesis, english.
16. Ziehe, A., Müller, K.R.: TDSEP – an efficient algorithm for blind separation using time structure. In Niklasson, L., Bodén, M., Ziemke, T., eds.: Proceedings of the 8th International Conference on Artificial Neural Networks, ICANN'98. Perspectives in Neural Computing, Berlin, Springer Verlag (1998) 675 – 680
17. Ziehe, A., Müller, K.R., Nolte, G., Mackert, B.M., Curio, G.: Artifact reduction in magnetoneurography based on time-delayed second-order correlations. IEEE Trans Biomed Eng. **47** (2000) 75–87
18. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer (1999)
19. Engelmann, D.M., Steitz, T.A., Goldmann, A.: Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu. Rev. Biophys. Chem. **115** (1986) 321–353
20. Chou, P., Fasman, G.: Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. **47** (1978) 45–148