ON THE PREDICTION OF BIOMINERALIZATION PROTEINS IN THE ABSENCE OF SEQUENCE HOMOLOGIES

Xiaoyu Zhang[1], Betsy Read[2]

[1]*Department of Computer Science,* [2]*Department of Biological Sciences*
California State University San Marcos
San Marcos, CA 92096

The major goal of the bioinformatics research at Cal State San Marcos is to identify gene and protein sequences of Coccolithophorids involved in biomineralization. However, most known biomineralization genes lack sequence homology, which makes it more challenging to identify such genes. On the other hand known biomineralization proteins share some common biochemical and biophysical characteristics: relatively small molecular weights, high acidity, repeats sequences, little or no secondary structure, propensity for trans-membrane helices, and high percentages of highly acidic amino acids (Asp, Glu, and Ser) in protein composition, etc. These features can be applied to predict candidate biomineralization proteins. We have developed a prediction scheme based on a statistical profile of features of known biomineralization proteins compared to randomly selected counterparts. Features in the predictor are combined and weights for the features are computed in order to minimize the relative entropy of wrong predictions in the training set. Given a new gene sequence, its features are calculated using tools from the ExPASy Proteomics Server. A biomineralization protein probability index is then computed by the predictor as a number between 0 and 1. We have tested the predictor on assembled EST and cDNA sequences of *Emiliania huxleyi* (E. Hux) and ranked them by their probability indices. The results reveal some interesting candidates whose involvement in biomineralization is currently being examined using microarray analysis and in-vitro biomineralization studies.

1. INTRODUCTION

Biomineralization is the process of forming hard tissues of organisms, e.g. shells, skeleton, and teeth etc, to fulfill a variety of different functions. The major goal of the bioinformatics research at Cal State San Marcos is to identify gene and protein sequences of Coccolithophorids involved in Bio-mineralization. However, most known biomineralization genes lack sequence homology, which makes it more challenging to identify such genes. Furthermore, more than 60% of assembled EST and cDNA sequences of Emiliania huxleyi, a model coccolithophorid organism, do not have significant hits to the known protein database. On the other hand known biomineralization proteins share some common biochemical and biophysical properties: relatively small molecular weights, high acidity, repeats sequences, little or no secondary structure, propensity for trans-membrane helices, and high percentages of highly acidic amino acids (Asp, Glu, and Ser) in protein composition, etc. We attempt to construct property profiles for known biomineralization proteins and apply them to predict biomineralization proteins from unknown gene sequences.

2. METHODOLOGY

The profiles of biomineralization proteins are built from 64 proteins that were downloaded from the NCBI protein database and verified in literatures to be related to biomineralization. A reference model is also constructed from 500 random protein sequences download from NCBI.

Following properties were collected for the current profile, most of which are computed using the ExPASy proteomics server[2]: Peptide length, Molecular weight and iso-electric point, Motifs found in the Prosite database[3], Domain families from ProDom search[1], Acidity, Percentages of Asp, Glu, and Ser in protein composition, Predicted phosphorylation sites, Predicted O-glycosylation sites[4], Predicted coiled-coil regions[5], Predicted trans-membrane helices[7], Predicted secondary structures[6], and Repeat sequences.

These properties are to be combined with results from sequence homology search. Additional properties may be added in the future work. A particular probability distribution is constructed for every property D_i, with different parameters for the set of biomineralization and random proteins. The set of probability distributions are the profiles to be used in the prediction step.

For a protein sequence with a certain value d for the property D_i, assume $P(D_i = d)$ is the probability from random protein profile and $P(D_i = d \mid B)$ is the conditional probability from the profile of

biomineralization proteins. According to Bayes theorem we have

$$\frac{P\,(\,B\,\mid\,D_i\,=\,d\,)}{P(B)}\,=\,\frac{P(D_i\,=\,d\,\mid\,B)}{P(D_i\,=\,d)}\,. \qquad (1)$$

Therefore $\frac{P(D_i=d|B)}{P(D_i=d)}$ provides the relative likelihood of B if $D_i\,=\,d$ is observed. We consider the logarithm of the relative likelihood and call it as the differentiation score of D_i:

$$r_i\,=\,\log\frac{P(D_i=d|B)}{P(D_i=d)}, \qquad (2)$$

Where $r_i\,>\,0$ if the protein is more likely to be related to biomineralization when $D_i\,=\,d$ is observed and $r_i\,<\,0$ otherwise. Not every property equally predicts biomineralization proteins. A weighted sum $x\,=\,\sum_i w_i\cdot r_i$ is used to compute the likelihood of the protein belonging to B, where w_i are weights selected for difference properties. This is a typical binomial classification problem with two classes, B and $\overline{B}$. The probability index of being a biomineralization protein is evaluated using the logistic function

$$y\,=\,\frac{1}{1+e^{-x}}\,. \qquad (3)$$

It is clear that $0\,<\,y\,<\,1$ and y increases for larger x. The weights w_i are determined using for a training set of proteins p_j, $j\,=\,1,\dots,N$ in order to minimize the relative entropy:

$$F\,=\,\sum_{j=1}^{N}-t_i\log y_i\,-\,(1-t_i)\log(1-y_i), \qquad (4)$$

where the target output t_j is 0 for non-biomineralization proteins or 1 for biomineralization proteins. The gradient of F can be computed as

$$\frac{\partial F}{\partial w_i}\,=\,\sum_j\frac{\partial F}{\partial y_j}\frac{\partial y_j}{\partial w_i}\,=\,\sum_j-(t_j-y_j)r_j^i\,. \qquad (5)$$

All weights are initialized to 0's and iteratively evaluated to find the minimum of F and corresponding w_i using the conjugate gradient method.

3. RESULTS AND DISCUSSIONS

Figure 1 shows the example distributions of acidity and Percentages of Asp, Glu, and Ser in protein composition for biomineralization and random proteins. If an unknown protein sequence has low PH value, it has a

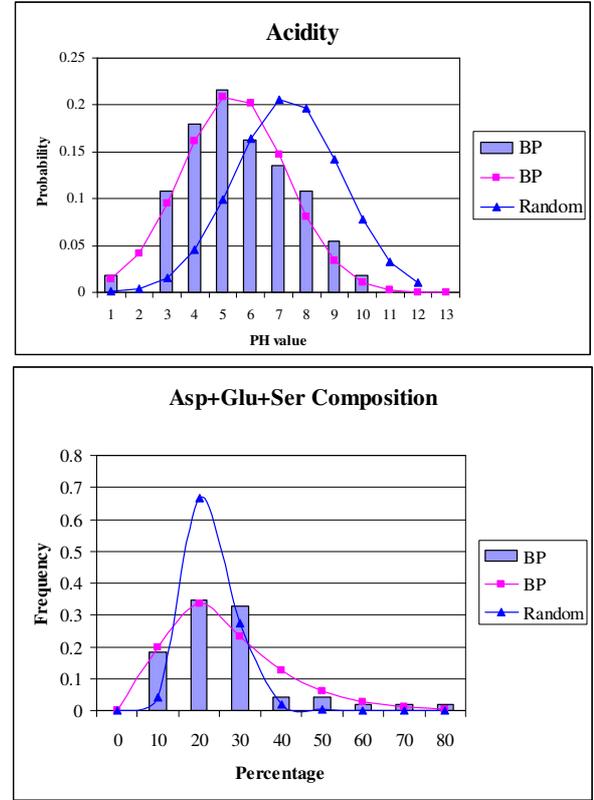positive differentiation score and is more likely a biomineralization protein.

Figure 1. Sample property distributions. The BP curves represent distributions of biomineralization proteins and Random curves for average random proteins.

We tested the prediction method on 190 full-length cDNA sequences of Emiliania huxleyi (E. Hux) and ranked them by their probability indices, part of which shown in table 1 (full list of results can be found at http://bioinfo.csusm.edu). The results reveal some interesting candidates whose involvement in biomineralization is currently being examined using microarray analysis and in-vitro biomineralization studies. The predictor is being refined to include more and updated properties in the profile and to be incorporated with results from homology search and micro-array analyses.

Table 1. Prediction probabilities on full-length cDNA sequences of Emiliania huxleyi

Sequence ID and reading frames	Probability	Motif	MW	acidicity(PI)	Asp+Glu+Ser	2nd Class	TMH	…
Plate1_C08_compl_77-2-8-A07_0	0.9633721	5	37452.7	3.96	30.80%	all-alpha	Y	…
Plate1_H04_compl_77-2-16-A09_rev_2	0.9297546	4	21602.04	6.73	31.90%	mixed	Y	…
Plate1_D11_compl_77-2-10-C04_rev_0	0.9175919	4	16762.72	7.56	35.00%	mixed	Y	…
Plate1_B09_3p_77-2-7-A09_0	0.8393985	4	5703.97	3.91	37.50%	mixed	N	…
Plate1_G02_compl_77-2-14-B10_rev_0	0.8365396	5	19042.19	10.78	30.90%	mixed	Y	…
….								
Plate1_B04_compl_77-2-5-F11_2	0.02916201	5	30810.96	8.24	17.70%	alpha-beta	N	…
Plate2_D10_compl_77-2-31-F12_rev_0	0.02728758	3	8447.7	11.55	16.90%	mixed	N	…
Plate2_G04_5p_77-2-38-H09_rev_0	0.02500642	4	10844.45	11.78	17.10%	mixed	N	…

Reference

[1] F. Corpet, J. Gouzy and D. Kahn, *The ProDom database of protein domain families*, Nucleic Acids Res, 26 (1998), pp. 323-6.

[2] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel and A. Bairoch, *ExPASy: The proteomics server for in-depth protein knowledge and analysis*, Nucleic Acids Res, 31 (2003), pp. 3784-8.

[3] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni and C. J. Sigrist, *The PROSITE database*, Nucleic Acids Res, 34 (2006), pp. D227-30.

[4] K. Julenius, A. Molgaard, R. Gupta and S. Brunak, *Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites*, Glycobiology, 15 (2005), pp. 153-64.

[5] A. Lupas, *Predicting coiled-coil regions in proteins*, Curr Opin Struct Biol, 7 (1997), pp. 388-93.

[6] B. Rost, G. Yachdav and J. Liu, *The PredictProtein server*, Nucleic Acids Res, 32 (2004), pp. W321-6.

[7] E. L. Sonnhammer, G. von Heijne and A. Krogh, *A hidden Markov model for predicting transmembrane helices in protein sequences*, Proc Int Conf Intell Syst Mol Biol, 6 (1998), pp. 175-82.