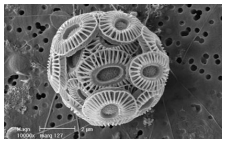
On the Prediction of Biomineralization Proteins in the Absence of Sequence Homologies

Xiaoyu Zhang[1], Betsy Read[2]
[1]Department of Computer Science, [2]Department of Biological Sciences
California State University San Marcos



ABSTRACT

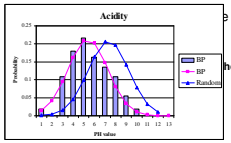
One major goal of the bioinformatics research at Cal State San Marcos is to identify gene and protein sequences of Coccolithophorids involved in biomineralization. However, most known biomineralization genes lack sequence homology, which makes it more challenging to identify such genes. On the other 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 incorrect 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* 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.

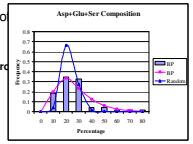
METHODS

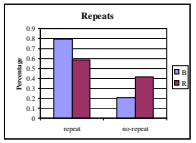
64 biomineralization proteins were downloaded from the NCBI protein database and verified in literatures to be related to biomineralization. A reference model was also constructed from 500 random protein sequences. To build a profile, we computed following properties for the biomineralization proteins and the reference model:

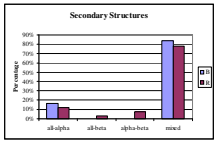
- Peptide length,
- Molecular weight and iso-electric point,
- Prosite Motifs [3],
- ProDom Domains[1],
- Acidity,
- Percentages of Asp, Glu, and Ser in protein composition,
- Phosphorylation sites,
- O-glycosylation sites[4],
- Coiled-coil regions[5],
- Trans-membrane helices[7],
- Secondary structures[6],
- Repeat sequences.

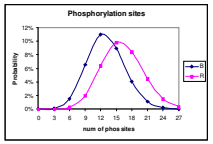
These properties are to be combined with results from sequence homology search. Additional properties may be added in the future. A particular probability distribution is constructed for every property, with different parameters for the biomineralization and random proteins. The

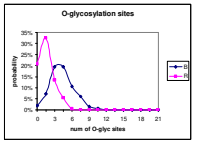


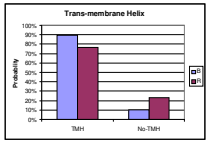


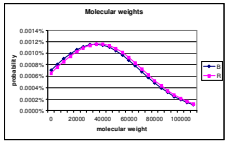












For a protein sequence with a certain value d for the property D, $P_R(D = d)$ is the probability for random protein profile while $P_B(D = d)$ the probability for the profile of biomineralization proteins. We consider the logarithm odds of the relative likelihood:

$$r = \log \frac{P_B(D=d)}{P_R(D=d)} \quad (1)$$

The odds $r > 0$ indicates the protein is more likely to be related to biomineralization when d is observed for the property D, conversely for $r < 0$. Because not every property equally predicts biomineralization proteins, a weighted sum is used to compute the overall odds of being a biomineralization protein,

$$x = \sum_i w_i \cdot r_i \quad (2)$$

where w_i are weights selected for difference properties.

The problem of predicting biomineralization proteins can be considered as a typical binomial classification problem with two classes. The probability index of being a biomineralization protein is evaluated using the logistic function

$$p = \frac{1}{1 + e^{-x}} \quad (3)$$

It is clear that $0 < p < 1$ and p increases for larger odds x.

The weights w_i in (2) are determined by using a training set of biomineralization and random proteins, in order to minimize the relative entropy:

$$F = \sum_{j=1}^{N} -t_j \log p_j - (1-t_j)\log(1-p_j) \quad (4)$$

where N is the total number of training proteins and the target output $t_j = 1$ if the j_{th} protein is related to biomineralization and $t_j = 0$ otherwise. The conjugate gradient method is used to solve for the minimum of F to get the appropriate weights w_i. The gradient of F relative to w_i can be computed as following,

$$\frac{\partial F}{\partial w_i} = \sum_{j=1}^{N} \frac{\partial F}{\partial p_j} \cdot \frac{\partial p_j}{\partial w_i} = \sum_{j=1}^{N} -(t_j - p_j) r_j^i \quad (5)$$

All weights are initialized to 0 and Equation (5) is iteratively evaluated until the minimum of F is reached.

RESULTS

We tested the prediction method on 190 full-length cDNA sequences of Emiliania huxleyi and ranked them by their probability indices. Parts of the results are 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 experiments.

Sequence ID and reading frames	Probability	Motif	MW	acidicity(PI)	Asp+Glu+Ser	2nd Class	TMH	…
77-2-8-A07_0	0.9633721	5	37452.7	3.96	30.80%	all-alpha	Y	…
77-2-16-A09_rev_2	0.9297546	4	21602.04	6.73	31.90%	mixed	Y	…
77-2-10-C04_rev_0	0.9175919	4	16762.72	7.56	35.00%	mixed	Y	…
77-2-7-A09_0	0.8393985	4	5703.97	3.91	37.50%	mixed	N	…
77-2-14-B10_rev_0	0.8365396	5	19042.19	10.78	30.90%	mixed	Y	…
			….					
77-2-5-F11_2	0.02916201	5	30810.96	8.24	17.70%	alpha-beta	N	…
77-2-31-F12_rev_0	0.02728758	3	8447.7	11.55	16.90%	mixed	N	…
77-2-38-H09_rev_0	0.02500642	4	10844.45	11.78	17.10%	mixed	N	…

Table 1. Prediction probabilities on full-length cDNA sequences of Emiliania huxleyi Sequence ID and reading frames

CONCLUSIONS

Since biomineralization proteins do not share sequence homology, we developed a feature-based prediction method for predicting such proteins. This method was applied to assembled EST and cDNA sequences of *Emiliania huxleyi*. The results reveal some interesting candidate biomineralization proteins that are currently under experimental 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.

REFERENCES

[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.