Bayesian Classifier for Anchored Protein Sorting Discovery

Fan Zhang
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, United States
zhangf@cse.sc.edu

Jianjun Hu
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, United States
jianjunh@cse.sc.edu

***Abstract*—A typical cell has a size of only 10 μm while it contains about a billion proteins. Transportation of these proteins from their synthesis sites to their target locations within or outside of the cell is precisely controlled by protein sorting signals. However, genome-wide understanding of protein sorting regulatory signals and mechanisms is still very limited. We formulate the protein sorting motif discovery problem as a classification problem and proposed a Bayesian classifier based motif discovery algorithm (BayesMotif) to find a common type of sorting motifs in which a highly conserved anchor is present along with a less conserved motif regions. Experiments showed that our algorithm has the advantage of finding long lowly conserved sorting signals compared to other protein motif discovery algorithms such as MEME. Our algorithm also has the advantage to easily include additional meta-sequence features that overcomes the limitation of PWM (position weight matrix)**

Keywords-protein sorting motif; motif discovery, Bayesian classifier; sorting signals

I. INTRODUCTION

A typical cell has a size of only 10 μm while it contains about a billion proteins. How these proteins are transported from their synthesis sites to their target locations within or outside of the cell is still not well understood. Experiments showed that translocation of nascent proteins are usually guided by "postal code" like targeting signals encoded within the amino acid sequences of proteins. Genome-wide identification and decoding of these molecular "zip codes" are fundamental to comprehensive understanding of the cell. Experimentally identifying protein targeting signals is labor and cost intensive, usually using a tedious cut-and-test approach [1;2]. Recently, genome scale protein localization data has become available [3] for a couple of species and gene ontology also provides a large amount of localization information of proteins [4]. These datasets provide a great opportunity for developing bioinformatic algorithms to identify protein sorting signals to guide biological experiments. However, computational prediction of targeting signals is still a big challenge due to their low conservation at the amino acid level. Many motif discovery algorithms [5] have been proposed in the past decades but mostly have been only tested or applicable to DNA motif discovery with alphabet of four nucleotides rather than 20 amino acids. Two commonly used de novo protein motif discovery algorithms are MEME [6] and TEIRESAS [7], which are not very effective to mine these signals due to the low conservation of sorting motifs at the amino acid level and/or their length.

In this paper, we are interested in de novo discovery of a common type of protein sorting motifs that are composed of a highly conserved anchor (2 to 5 amino acids long) and a less conserved amino acid region with a specific physichemical property. Most of these sorting signals are located within the 200 amino acids of the N-terminal or C-terminal. For example, Chaddock et al. [8] examined thylakoid transfer signals from all of the known lumenal proteins and found that all of the substrates for the ApH-dependent translocase possess a twin-arginine motif (RR) immediately before the hydrophobic (H) amino acid region. Brink [9] showed that the RR motif alone is not sufficient for the delta pH transportation and another signal inside the hydrophobic region is required. Sheikh and Isacke reported a di-hydrophobic motif Leu330-Val334 motif which is located within a cytoploasmic domain [10].

II. METHODOLOGY: BAYESIAN CLASSIFIER FOR PROTEIN SORTING MOTIF DISCOVERY

A. Overview of the algorithm

We formulate the protein sorting motif discovery problem as a classification problem: Given a set of protein sequences $P = \{s_1, s_2, ...s_N\}$ that are localized to the same location L, a negative set N of sequences are selected composed of proteins that are not localized to location L. Identification of sorting motifs can be thus mapped to finding a motif model which can differentiate the motif instances from positive sequence set from background of the negative sequences. The higher the classification accuracy of a motif model has to differentiate positive sets from negative sets, the better the motif model.

We are interested in protein sorting motifs that are composed of a highly conserved, but short anchor (mostly these anchors have fewer amino acids than 4, e.g.: in RR translocation pathway, the signal peptide all have a twin-arginine pair located between N and H region, and for LDL receptors, an NPXY motif frequently shows up at COOH terminal of the sequence) and a comparably low-conserved motif region around the anchor. Because most of the sorting motifs are not well conserved at the amino acid level, it will be difficult to find out these motifs by sequence alignment. Our approach is to firstly search the most frequent anchors in positive datasets, then use

Bayesian classifier to test if an anchor has a motif region around which can well differentiate them from background sequences (negative dataset). Our method is also able to determine motif boundary by a sliding-window test on cross validation score returned by Bayesian classifier.

Our motif finding algorithm is composed of three major steps (Figure 1):

1) Preprocessing protein sequences by cutting protein sequences and keep only N-terminal and C-terminal K amino acids and then do redundancy removal
2) Finding frequent anchors by regular expression enumeration;
3) Constructing Bayesian classifier to detect low conserved motif regions around anchors;
4) Based on the motif boundary given by step 3, calculate discrimination score for each motif using cross-validation test on Bayesian classifier again.

B. Preprocessing of datasets

In protein sorting motif discovery problem, a given set of proteins assumed to be transported to a specific location are given. These proteins can be either obtained from gene ontology annotation or genome scale localization experiments. For each such sequence, we will cut 200 amino acids from the N-terminal and C-terminal and apply the motif discovery algorithms on them.

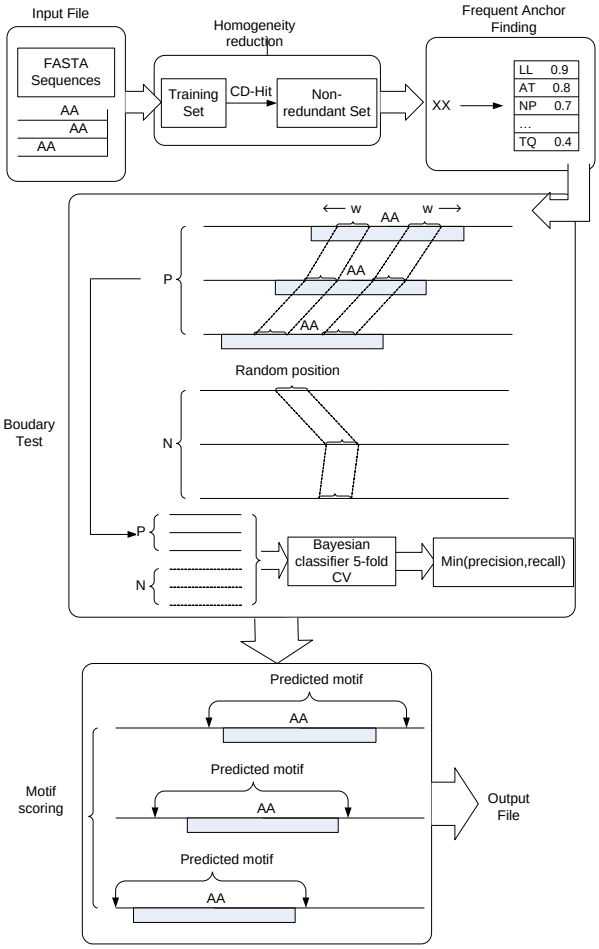


Figure 1. BayesMotif discovery algorithm

To evaluate our algorithm, we use both synthetic datasets and real datasets from Swiss-Prot release 48; synthetic dataset are generated by adding an artificial motif randomly in a set of protein sequences: firstly choose an arbitrary set of protein sequences (we use cytoplasmic protein in our experiments), and then divide the whole set into balanced two partitions. One of the partitions is used as negative training samples. An artificial motif is inserted into a random position in the N-terminal of each sequence in the positive set. The random motifs are created with a centered 2-amino acids anchor, and a surrounding region composed of amino acids drawn from prior probabilistic distributions, which is used to simulate the hydrophobic, charged or other physichemical regions typically occurring with the anchors.

The real dataset we use are two protein datasets contain real motifs: RR translocation signal peptide and LDL receptors, a background protein set extracted from cytoplasmic localization are used as negative training samples to real positive dataset. Because the number of positive training samples are comparably fewer than negative, we randomly select a subset from all cytoplasmic proteins from Swiss-Prot database to make the two classes to be more balanced, also considering the possible noises and duplications in data, we use cd-hit, a sequence cluster and homogeneity reduction tool, to reduce the homogeneity in both positive and negative training samples, the similarity threshold is set up to be 80%.

An important preprocessing step is to remove redundancy in the training sequences. The rationale is that redundancy in training samples will lead to classifiers biased to the over-represented class composed of redundant training samples, which therefore leads to misleading prediction accuracy. To reduce the redundancy in the dataset, we use CD-HIT [11], a sequence clustering algorithm. CD-HIT has the ability to cluster the sequences by predefined or user defined weight matrices and a similarity threshold, and remove all identical sequences in the same cluster but the pivot, which guarantees each pair of sequences in those left pivots will not be similar to each other, to make sure all sequences are not identical globally, the threshold for redundancy reduction in the experiments is set to 80%

C. Frequent Anchor Discovery

Frequent anchors are identified using exhaustive regular expression searching on positive dataset. The search space is defined on a gap-tolerant regular expression anchor model since many protein sorting motifs (e.g. NPXY and YXXϕ motif in LDL receptors) are not completely conserved amino acid sequences, but a combination of two motifs with a variable-length gap. To find out these more flexible anchors, we use a regular expression model with the form: <Amino Acid>{n}<X><Amino Acid>{m} to represent the "language" of possible anchors: the anchor model is composed of two motif region <Amino Acid>{n} and <Amino Acid>{m}, which is two informative regions of amino acids with length n and m, and a gap toleration

between the two motif regions <X> {min, max}, min and max are two parameters to control gap length, the length of the gap must be in the interval defined by {min, max}, to be more adaptive, we allow the two motif region also have controllable length and allow them to have different amino acid alphabets. Using this regular expression model, we can then enumerate all possible anchors and count their occurrence frequency in the positive dataset in both N and C terminal regions. We then check if there are conserved regions around these anchors and how these regions can differentiates positive datasets from negative ones.

D. Motif Boundary Determination

After generating the ranked anchor list, Bayesian classifiers are trained to identify the most likely boundary of low-conserved motif region around the anchors. For each anchor occurrence at N or C terminal in positive dataset, the algorithm use a window of fixed length W to slide from to the left and right of the anchor, each time using the amino acids in the window for all positive sequences as input to train a Bayesian classifier. For negative dataset, a randomly picked window within N or C terminals is used for extracting background samples for training. After training a Bayesian classifier, we use 5-cross validation to obtain the prediction accuracy of the classifier for a given sliding window. If the smaller value of precision and recall is lower than a threshold score (e.g. 0.5), it means the sliding window is moving out of the true motif region and the left boundary can thus be determined. It is obvious that the farther the sliding window leaves the motif, the more irrelevant regions will be included in the window, so the lower the score will be. Similarly, the right boundary can be decided.

E. Scoring Motif Discrimination Capability and Conservation

After left and right boundaries for each anchor are determined, we pick up the sub sequences between left and right boundaries, and train a Bayesian classifier again to get the overall classification score of the motif region, which reflects the capability of the motif to differentiate positive proteins from negative datasets or proteins targeting a specific subcellular localization against proteins that target elsewhere. The score is defined as min (precision, recall).

Measuring Motif Conservation

To measure the conservation of discovered motifs, we use the information content measurement: for a fixed length motif model, each position in the model can be seemed as a random variable, the entropy of this random variable can be calculated by Shannon theorem, let S be the sequence set of a motif, s_i be the ith sequence in S, p_j be the jth position in the sequences, $\sum$ be the amino acid alphabet.

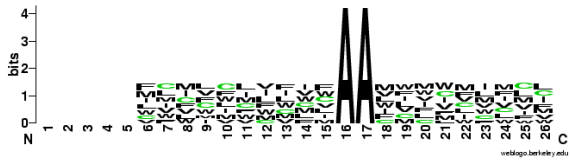
$$I(p_j) = \sum_{a \in \Sigma} p(a) \log_2 p(a)$$

$$I(S) = \sum_{p_j \in S} I(p_j)$$

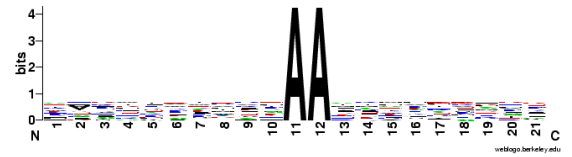
where $p(a)$ can be calculated by counting the frequency of different amino acids appeared in different positions in the sequence set, $I(s)$ is the joint information of all positions in the motif. In our case, it will be the information content of the motif.

III. EXPERIMENTAL SETUP

The simulation dataset are prepared as follows: First we randomly select 2000 cytoplasmic protein sequence set from Swiss Prot release 48 database. 1000 of them are used for constructing positive dataset, and 1000 as negative dataset. For each positive sequence, an artificial anchor (in our experiments are ----AA----) is first inserted into a random position of the first 100 N-terminal amino acids. And then two equal-sized regions are inserted around the anchor. These two are low-conserved and are generated according to a prior distribution on amino acids, e.g. hydrophobic amino acids. We use cytoplasmic proteins for both positive and negative datasets in order to guarantee they share identical background distribution of amino acids. Three sets of motifs are implanted into three positive datasets (Figure 2.). For real datasets, Tat-pathway translocation proteins and LDL receptors are extracted from Swiss Prot as shown in Table I.



a): simulated hydrophobic motifs anchored by AA



b): simulated hydrophobic/positively charged motifs anchored by AA

c) No conserved motifs around anchors

Figure 2. Logos of implanted motifs with a fully conserved anchor and less conserved physichemical motifs.

TABLE I SIMULATED AND REAL DATASETS

Dataset	Number of Positive samples number	Number of Negative samples number	Anchors
Synthetic	1000	1000	----AA---- (Artificial)
Tat-Pathway Translocation	86	600	----RR----
LDL receptor	464	439	----NPXY----

IV. RESULTS

A. *Results on Synthetic Datasets*

In our experiments, we simulate the low-conserved motif region with amino acids on hydrophobicity index and charged property, tests on the datasets show that our method has a strong capability to distinguish boundary of anchored lowly-conserved artificial motifs.

TABLE II: BOUNDARY TEST RESULTS FOR IMPLANTED MOTIF MODELS

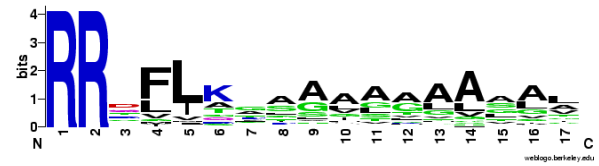
Motif	Implanted Motif length	Detected motif length	Motif Entropy	Motif Score
Hydrophobic	20	25	73.8	100
Hydrophobic+ Charged	20	21	67.7	99.7
Random	20	8	26	57.6

First two rows of Table II show that when artificial motifs are created in a information-rich distribution, the boundary test can always find the right motif region, in comparison of hydrophobic and charged motif, the third row shows when the motif are generated in a totally random way, the boundary algorithm returned is no longer accurate compared with actual motif boundary, which make sense because it will be difficult to give boundary prediction if the motif is of no difference from the background, also we can see from column 5 and 6, the entropy per amino acid and discriminate score are inversely correlative, high discriminate score corresponds to low uncertainty in amino acid distribution thus a low value in information content will be returned and vice versa.

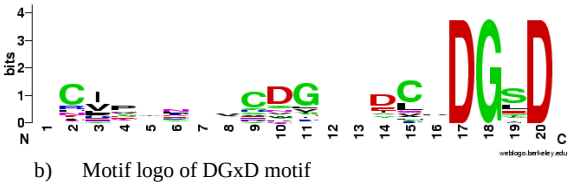
B. *Results on real datasets*

1) De-novo discovery of RR translocation signal peptide RR-x-FLK

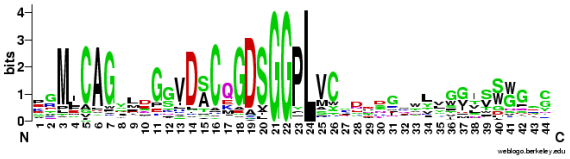
TAT system is known as Sec-independent protein export pathway in bacteria. The most remarkable feature in TAT translocation proteins is the presence of the double arginines located between N and H region of the signal peptide. We downloaded 86 Tat-translocation proteins from SwissProt database and applied our BayesMotif algorithm with a two-amino acid XX anchor model. A set of 600 cytoplasmic proteins are used as negative dataset. After homogeneity reduction with CD-Hit, our BayesMotif algorithm found the following motif with 17 amino acids. The motif score is 87. 9, which means that the classifier can achieve classification accuracy of at least 0.879 in precision or recall rate. Although a functional RR-consensus motif RR-X-FLK is indispensable for targeting the Tat translocase, additional sequence features of RR-signal sequences seem to be required to prevent mistargeting to the Sec export pathway [12].



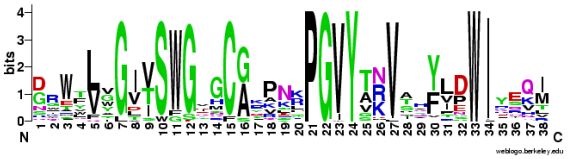
a) Motif logo of TAT-Translocation signal peptide RRxFLK



b) Motif logo of DGxD motif



c) Motif log of GGPL and GDSG motif



d) Motif log of putative Motif PGVY

Figure 3: discovered sorting motifs

2) De-novo discovery of NPxY motif at C terminal of Megalin LDL receptor

Megalin is the main endocytic receptor of the proximal tubule and is responsible for reabsorption of many filtered proteins. It is found that information that directs apical sorting is present in the cytoplasmic tail (CT) of megalin, which contains three NPXY motifs, YXXØ, SH3, and dileucine motifs, and a PDZ-binding motif at its COOH terminus. Using 464 megalin sequences downloaded from Swiss-prot database as positive dataset and cytoplasmic proteins as negative dataset, BayesMotif algorithm found the NPxY motif at the C-terminal along with a conserved amino acid region with undiscovered biological functionality.

Besides the NPxY motif, we also found two other biologically verified motifs: DGxD motif and GGPL motif. DGxG motif is found in the alignment of five ligand-binding repeats in rat LRP3 in comparison with a consensus sequence of those in LDL receptors, a C-terminal DGSDE pentapeptide, which forms part of the ligand-binding site of LDL receptors and is almost completely conserved. GGPL motif not only appears in LDL receptors but also in other protein families as GRF1-4 and OsGRF1, which presents as a C-terminal motif essentially related to transactivition activity [13].

We also noticed two motifs with a significant high score found by our algorithm: GDSG and PGVY motifs. GDSG motif (Figure 2.a) has a long motif region overlapped with GGPL motif, implying that it could work as a functional part of GGPL motif. PGVY is a new independent motif which has a well conserved motif

region. The biological interpretation of this motif is still unknown yet, but significance from both frequency counting, sequence entropy and discrimination scoring suggests that the over representation of this motif is not likely to be coming from randomness of amino acid combination in proteins but has some biological significance.

C. Comparison with other motif algorithms

We also input the same datasets to two other popular protein motif discovery algorithms: MEME and Teiresias [7]. MEME use Position Weighted Matrix as motif model and search overrepresented patterns on a given dataset by maximizing the motif likelihood using an EM algorithm. Teiresias is a regular expression modeling algorithm for motif finding. Teiresias adapts apriori method for frequent pattern mining to solve protein motif discovery problem. Even though our BayesMotif uses a supervised classification algorithm for motif search, it is essentially a de novo motif discovery algorithm comparable to MEME.

We have tested these two algorithms on the simulated datasets. It turned out that Teiresias cannot retrieve any of the implanted motifs due to its inability to find long motifs. MEME can find the implanted motifs but reported them as two separate motifs. We then tested the two algorithms on the real datasets and we found that MEME and Teiresias can identify the following motifs RR-FLK, GGPL, and PGVY. But MEME failed to find the NPxY, GDSG, and DGxD motifs while Teiresias failed to find NPxP and DGxD motifs. However, both MEME and Teiresia tend to find short motifs while most protein sorting signals are composed of a short anchor and a region with low-conservation, which poses difficulty for conventional algorithms.

V. CONCLUSIONS

We proposed a Bayesian classifier based protein motif discovery algorithm for de novo identification of anchored protein sorting motifs. Experiments on both simulated datasets and real datasets demonstrated that the proposed BayesMotif algorithm is able to retrieve implanted motifs as well as experimentally identified biological motifs. Compared to conventional motif discovery algorithms, the classification algorithm formulation of BayesMotif makes it easy to incorporate additional structural or meta-sequence features for motif discovery such as hydrophobic or secondary structures and etc. Another advantage is that the BayesMotif algorithm can work on very large datasets while current algorithms may not handle. It should be noted that the positive dataset can be easily picked by identifying a set of proteins that target to the same subcellular location. The negative dataset simply include proteins that do not target to that location.

ACKNOWLEDGMENT

This work is partially supported by NSF under grant 0845381. We appreciate the discussion with other members of the Machine Learning and Evolutionary Laboratory at USC. The web server for this program will be made available at http://mleg.cse.sc.edu/sortmotif.

REFERENCES

[1] A.M.Chaddock, A.Mant, I.Karnauchov, S.Brink, R.G.Herrmann, R.B.Klosgen, and C.Robinson, A new type of signal peptide: central role of a twin-arginine motif in transfer signals for the delta pH-dependent thylakoidal protein translocase. EMBO J. 14 (1995) 2715-2722.

[2] W.J.Chen, J.L.Goldstein, and M.S.Brown, NPXY, a sequence often found in cytoplasmic tails, is required for coated pit-mediated internalization of the low density lipoprotein receptor. J.Biol.Chem. 265 (1990) 3116-3123.

[3] J.L.Fink, R.N.Aturaliya, M.J.Davis, F.S.Zhang, K.Hanson, M.S.Teasdale, C.Kai, J.Kawai, P.Carninci, Y.Hayashizaki, and R.D.Teasdale, LOCATE: a mouse protein subcellular localization database. Nucleic Acids Research 34 (2006) D213-D217.

[4] M.A.Harris, J.Clark, A.Ireland, J.Lomax, M.Ashburner, R.Foulger, K.Eilbeck, S.Lewis, B.Marshall, C.Mungall, J.Richter, G.M.Rubin, J.A.Blake, C.Bult, M.Dolan, H.Drabkin, J.T.Eppig, D.P.Hill, L.Ni, M.Ringwald, R.Balakrishnan, J.M.Cherry, K.R.Christie, M.C.Costanzo, S.S.Dwight, S.Engel, D.G.Fisk, J.E.Hirschman, E.L.Hong, R.S.Nash, A.Sethuraman, C.L.Theesfeld, D.Botstein, K.Dolinski, B.Feierbach, T.Berardini, S.Mundodi, S.Y.Rhee, R.Apweiler, D.Barrell, E.Camon, E.Dimmer, V.Lee, R.Chisholm, P.Gaudet, W.Kibbe, R.Kishore, E.M.Schwarz, P.Sternberg, M.Gwinn, L.Hannick, J.Wortman, M.Berriman, V.Wood, C.N.de la, P.Tonellato, P.Jaiswal, T.Seigfried, and R.White, The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32 (2004) D258-D261.

[5] W.Wei and X.D.Yu, Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. Genomics Proteomics Bioinformatics 5 (2007) 131-142.

[6] T.L.Bailey, N.Williams, C.Misleh, and W.W.Li, MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 34 (2006) W369-W373.

[7] I.Rigoutsos and A.Floratos, Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics 14 (1998) 55-67.

[8] A.M.Chaddock, A.Mant, I.Karnauchov, S.Brink, R.G.Herrmann, R.B.Klosgen, and C.Robinson, A new type of signal peptide: central role of a twin-arginine motif in transfer signals for the delta pH-dependent thylakoidal protein translocase. EMBO J. 14 (1995) 2715-2722.

[9] S.Brink, E.G.Bogsch, W.R.Edwards, P.J.Hynds, and C.Robinson, Targeting of thylakoid proteins by the delta pH-driven twin-arginine translocation pathway requires a specific signal in the hydrophobic domain in conjunction with the twin-arginine motif. FEBS Lett. 434 (1998) 425-430.

[10] H.Sheikh and C.M.Isacke, A di-hydrophobic Leu-Val motif regulates the basolateral localization of CD44 in polarized Madin-Darby canine kidney epithelial cells. J.Biol.Chem. 271 (1996) 12185-12190.

[11] W.Li and A.Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22 (2006) 1658-1659.

[12] M.Muller, Twin-arginine-specific protein export in Escherichia coli. Res.Microbiol. 156 (2005) 131-136.

[13] J.L.Ditty and C.S.Harwood, Charged amino acids conserved in the aromatic acid/H+ symporter family of permeases are required for 4-hydroxybenzoate transport by PcaK from Pseudomonas putida. J.Bacteriol. 184 (2002) 1444-1448.