Poster H-26

Evolution of the Inorganic Pyrophosphatase Family

Authors:
Joel Hedlund (*IFM Bioinformatics, Linköping University, Linköping, Sweden*)
Roberto Cantoni (*CGB, Karolinska Institute, Stockholm, Sweden; Department of Physical Sciences, "Federico II" Univers*)
Margareta Baltscheffsky (*Arrhenius Laboratories, Stockholm University, Stockholm, Sweden*)
Herrick Baltscheffsky (*Arrhenius Laboratories, Stockholm University, Stockholm, Sweden*)
Bengt Persson (*IFM Bioinformatics, Linköping University, Linköping, Sweden; CGB, Karolinska Institute, Stockholm, S*)

Short Abstract: Inorganic pyrophosphatases contain ancient sequence motifs and produce energy through catabolism of pyrophosphate molecules, a mechanism thought to predate ATP driven cellular energy production. We have used bioinformatic methods to discover new family members, to study motifs and evolutionary relationships, and to identify potentially determining features for differences among subfamilies.

Long Abstract:
Inorganic pyrophosphatases (H+-PPases) form a strongly conserved family of tightly membrane bound proteins generally found in plants, prokarya and some protozoa, and whose main function is energy production through catabolism of pyrophosphate molecules. The family is interesting for evolutionary studies since it contains several well conserved ancient sequence motifs, and also because its function is thought to be the predecessor to ATP driven cellular energy production. In this work we have used bioinformatic methods to identify distantly related and previously unknown members of this family. We have also studied sequence motifs typical of this family and its evolutionary relationships. Furthermore, we have identified several sites potentially determining for functional differences among subfamilies. We show that for the inorganic pyrophosphatase family, sequence motifs, functional residues and conserved residues and regions are all located in cytosolic loops and transmembrane regions.

Methods:
Pyrophosphatase sequences were collected from Uniprot and genomeLKPG (in-house maintained database of all genomes in the public domain) using a HMMer hidden Markov model based upon the highly conserved 57-residue segment between transmembrane segments 5 and 6. In total, 145 sequences were found. Remarkably, only other H+-PPases were found using this strategy, indicating high specificity of the model. Dialign-t was used to build a multiple sequence alignment. Sequence fragments were excluded in order to enhance multiple sequence alignment quality, and sequences more than 99% identical were also removed. Residue and windowed regional conservation were then calculated from the resulting multiple sequence alignment.

ClustalX was used to build a phylogenetic tree, in which two distinct subgroups could be identified, corresponding to type 1 and type 2 H+-PPases. The multiple sequence alignment

was then processed by in-house produced software to yield information on functionally significant substitutions. In this algorithm the conservation rates for both subgroups are calculated separately for each position in the multiple sequence alignment by summing the relative frequencies for the two most common amino acid residue types in the respective subgroups. If the combined relative frequencies both are greater than 95%, and the most common amino acid residue types differ between the groups, the position is regarded as having a significant substitution. In case the dominating residue types have dissimilar properties, then the substitution is regarded as having a potential functional impact. Transmembrane topology was inferred from experiments on HPPA_STRCO. The substitution and conservation information along with known sequence motifs were mapped onto the transmembrane topology prediction using our own software, which also compiles printable topology plots (shown in the poster).

Some plant, archaeal, alvelolatean and euglenozoan species have sequences present in both subgroups. These sequences were extracted and analyzed species wise for adaptive evolution, both windowed and globally, using our own implementation of the Nei-Gojobori algorithm.

Results and discussion:
The initial search for new protein family members yielded more than twice as many inorganic pyrophosphatase sequences than were currently known. The hidden Markov model was also used to search the 1 million-odd sequences recently reported from a large-scale sequence project of organisms in the Sargasso Sea resulting in additional 168 partial pyrophosphatase sequences. The 57-residue cytosolic loop between transmembrane segments 5 and 6 seems to be unique for the H+-PPase family, making it suitable as a fingerprint in the search for further members of this family from new sequence data. This 57-residue region contains two nonapeptide sequences mainly consisting of the four "very early" proteinaceous amino acid residues Gly, Ala, Val and Asp, compatible with an ancient origin of the inorganic pyrophosphatases. The nonapeptide patterns have charged amino acid residues at positions 1, 5 and 9, and are apparent binding sites for the substrate and parts of the active site.

From the transmembrane mapping it is evident that known sequence motifs, functional residues and conserved residues and regions are all located in cytosolic loops or transmembrane regions in the inorganic pyrophosphatase family. In total, 22 such differences are found: 13 in the transmembrane segments and 9 on the cytosolic side, while none are found on the non-cytosolic side. Three differences could be expected to have functional impact as judged from the residue type exchanges. The implied functional impact has already been confirmed for one of these differences (position 507 in HPPA_STRCO), as an Ala/Lys mutation introduced at the corresponding position in Carboxydothermus hydrogenoformans type 1 H+-PPase has by others been shown to confer the potassium independence of type 2 H+-PPases to the enzyme. Transmembrane substitutions with implied functional impact are mainly concentrated to transmembrane segments 6, 8, 13 and 15, which implies that these segments may be closely located in tertiary structure. One of the significant eukaryotic substitutions is located in a sequence motif that may be conserved from a very early point in evolution (DNVGDNVGD, further details in poster).

No kind of adaptive evolution was detectable in any of the sequences, which implies that their respective branches have been functionally stable for more than 100 million years.

For a number of plant H+-PPases, the hidden Markov model finds distant homologies also to a second region of the enzymes, possibly being visible traces of an ancient gene duplication. This second region is located at residues 738–785 (numbering according to the A. thaliana sequence with accession number Q9FWR2). The patterns of this second region are also seen in further species variants, but are most clearly distinguishable in the plant sequences.