An Overview of Proteome Analyst

Roman Eisner, Brett Poulin, Duane Szafron, Paul Lu, Russell Greiner, David Wishart and Bahram Habibi-Nazhad
University of Alberta
www.cs.ualberta.ca/~bioinfo/PA

Proteome Analyst has received Significant funding from PENCE

new Proteome

> unknown 1
MRVLKFGGT...
> unknown 2
MVKVYAPAS...
> unknown 3
MPDFFSFIN...
> unknown 4
MPDFFSFIN...
> unknown 5
MTDKLTSLR...
> unknown 6
MGKIIGIDL...
> unknown 7
MSMSHINYN...
.....

What if I have cDNA instead of proteins

What about sequences with no similarity matches?

Proteome Analyst

SwissProt Database

Psiblast

unknown sequences with no similarity matches

spIP00561IAK1H_ECOLI BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 1114 0.0
spIP27725IAK1H_SERMA BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 1074 0.0
spIP49079IAKH1_MAIZE BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 1023 0.0
spIP49080IAKH2_MAIZE BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 993 0.0
spIP44505IAKH_HAEIN BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D... 981 0.0
spIP37142IAKH_DAUCA BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D... 980 0.0
spIP57290IAKH_BUCAI BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D... 974 0.0
spIP00562IAK2H_ECOLI BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 818 0.0

similar sequences for unknown-1

unknown sequences that have high similarity to some sequences in the SwissProt database

most similar sequences for unknown-1

LOCUS AK1H_ECOLI 820 aa linear BCT 16-OCT-2001
DEFINITION Bifunctional aspartokinase/homoserine dehydrogenase I (AKI-HDI) [Includes: Aspartokinase I; Homoserine dehydrogenase I].
ACCESSION P00561
PID g113539
VERSION P00561 GI:113539
DBSOURCE swissprot: locus AK1H_ECOLI, accession P00561; class: standard. extra accessions:Q47659,created: Jul 21, 1986. sequence updated: Jul 21, 1986. annotation updated: Oct 16, 2001.
xrefs (non-sequence databases): EcoGene EG10998, InterPro IPR002912, InterPro IPR001048, InterPro IPR001341, InterPro IPR001342, Pfam PF00696, Pfam PF01842, Pfam PF00742, PROSITE PS00324, PROSITE PS01042
KEYWORDS Transferase; Kinase; Oxidoreductase; Threonine biosynthesis; NADP; Allosteric enzyme; Multifunctional enzyme; Complete proteome.
SOURCE Escherichia coli.

Feature extractor

Transferase; Kinase; Oxidoreductase; Threonine biosynthesis; NADP; Allosteric enzyme; Multifunctional enzyme; Complete proteome. IPR002912; IPR001048; IPR001341; IPR001342. Transferase; Kinase; Oxidoreductase; Threonine biosynthesis; NADP; Allosteric enzyme; Multifunctional enzyme. IPR002912; IPR001048; IPR001341; IPR001342. Transferase; Kinase; Oxidoreductase; Methionine biosynthesis; NADP; Multifunctional enzyme; Chloroplast; Transit peptide; Multigene family. IPR002912; IPR001048; IPR001341.

features

General Classifier

Amino acid biosynthesis

How does a classifier work?

Proteome Analyst Card — **Unknown - 1**

Last_Update	May 7, 2002
Name	unknown 1
Alternate_Names	1) Homoserine dehydrogenase 2) L-homoserine: NAD(p) + oxidoreductase
General_Function	Amino acid biosynthesis and metabolism
Specific_Function	Aspartate metabolism; Glycine metabolism; Serine metabolism; Threonine biosynthesis
Gene_Ontology	Function: # No Data; Cell Process: # No Data; Cell Component: # No Data
Pfam_Domain/Function	PF01842 ACT domain 316-384; PF00742 Homoserine dehydrogenase; PF00696 Amino acid kinase family
Similarity	In the N-terminal section; belongs to the aspartokinase family. In the C -terminal section; belongs to the homoserine dehydrogenase family
Cell_Location	Cytoplasm
EC_Number	1.1.1.3
#_Amino_Acids_T	820 (Translated Protein)
#_Amino_Acids_M	819 (Mature Protein)
Calculated_Mw_(Daltons)_T	89122.9 (Translated Protein)
Calculated_Mw_(Daltons)_M	88991.7 (Mature Protein)
Theoretical_pI_T	5.63 (Translated Protein)
Theoretical_pI_M	5.43 (Mature Protein)
Observed_pI	unknown

How does proteome analyst predict other annotation information in the case where similar sequences are found?

Low Similarity Sequences! What do we do?

Analysis techniques are being developed based on motifs

Motif finder

Secondary Structure P.

cDNA?

No problem.
Proteome Analyst will pre-translate to protein.

> 2.A5.73
5' End, Reading Frame = 3
CCCTGACAGTCGACACTAGTGGATCCAAGAATTCGGCACGAGGCGAGCACGGACGGGTTCCTGC

GTCAACAGTGCTTGGACGGAACCGGCGCGCTCGTCGTCGACCGCCGTCACCACCACCACGACGCACGCGA

CACCGCCGCCTCCATCTCAATCTCATCCAGCGCCATGGCTACGCCTCCTCCCAGGTGCGCAGAACTACC
M A T P L S Q V R Q N Y
ACCAGGACTGCGAGGCGCGTCAACCGCCAGATCAACCTGGAGCTCTACGGTCTCTACGTGTACCTCAGCA
N Q D C E A A V N R Q I N L E L Y A S Y V Y L S
TGTCCTACTTATTGACCGGATGTATGTGCTCTGAAAACTTTGCCAAGTACTTCCTGCACGACGTCCACG
M S Y Y F D R D D V A L K N F A K Y F L H Q S H
AGCAACGGGAGCAGCCTGAGAAGCTGATGAAACTGCAAATCGAGGAGCGGCCGCATCTTCTTGCAGCACA
E E R E H A E K L M K L Q N Q R G G R I F L Q D
TCAAGAAACCAGATCGTGATGACTGGGAGAATGGCGTGACTGCAATGGAGTGTGCCGTGCACCTGGAGAAGA
I K K P D A D D W E N G L T A N E C A L H L E K
ACGTGACCGACGTCGCTGTCGAGCTGCACAAATTGGCAACTGAAAGAACGACCCACTCTGTTGACTTCA
N V N Q S L L E L H K L A T E K N D P H L C D F
TCGAACTCACTACCTGCGATGACAGGTGAAGCCATCAAGGACGCTCGCGACCATCTCACCAACCTGCGGA
I E T H Y L D E Q V K A I K E L G D H V T N L R
AGATGGGGGCACCCAAGTATTGGCAATTGGCAAGAGTACCCTTTGACAAGCACGCCCTGGGGCACAGTG
K M G G T Q V L A L A R V P L
ATTGAGAAGCTTGAAGTCTTGCCCACGAAACAACCACCCTGGTGGGTTTTCAAGGAAGAATTCCTTA

CTTTTAACTTGGTCCAAGCAAGTGCATTGCATTCTTCCNGCNTATTCTAAAATTAAACAAGGTTCTTCTTT

AACTCCTTGGTAACCAAAATATTCAN

Creating a Custom Classifier

training (example) sequences in fastA format with the correct class name first

>Kv1<VIC0..
mtvatgdpa...
>Kv1<VIC2..
mtvatgdpv...
>Kv2<VIC68..
mrseksltl...
>Kv3<VIC125.
mlssvcvss...
...

Proteome Analyst

custom classifier

Proteome Analyst evaluates the quality of each custom classifier using k-fold validation

Table Of Classification Validation Results:

True Class \ PA Predicted	A	B	C	D
A	0.955	0.000	0.000	0.045
B	0.167	0.833	0.000	0.000
C	0.167	0.000	0.833	0.000
D	0.056	0.222	0.000	0.722

LEGEND: A Kv1, B Kv2, C Kv3, D Kv4

Evaluation of a custom K-ion channel classifier with a 70 sequence training set

Classifiers

Classifier Algorithms (Machine Learning)

- Naive Bayes
- Tree-Augmented Naive Bayes (TAN)
- Neural Networks
- Decision Trees
- Support Vector Machine (SVM)

- Feature Selection - Wrapping

Classifier Decisions

> unknown 13
MKKLQIAVG...

Feature extractor

Mutator protein; DNA replication; DNA repair; Hydrolase; Magnesium; IPR000086;

How do I make my own custom classifier?

General Classifier

Histogram of Classification Probabilities:

Probability

Classification

Legend:
A: Cellular processes
B: Regulatory functions
C: Fatty acid and phospholipid metabolism
D: Replication
E: Biosynthesis of cofactors
F: Other categories
G: Purines
H: Transport and binding proteins
I: Translation
J: Energy metabolism
K: Cell envelope
L: Amino acid biosynthesis
M: Central intermediary metabolism
N: Transcription

Explain this decision

Replication 67%

Classifier Decisions Explained

Class Prediction
Magnified Contributions of Selected Tokens

Other categories
Purines
Energy metabolism
Biosynthesis of cofactors
Amino acid biosynthesis
Central intermediary metabolism
Translation
Cell envelope
Transport and binding proteins
Cellular processes
Replication
Transcription
Fatty acid and phospholipid metabolism
Regulatory functions

Normalized Sum of Conditional Probabilities (Logarithmic)

dna repair, dna replication, ipr000086, magnesium, hydrolase, Residual, Remaining Tokens

The features (tokens) in the sequences that were most important in making the classification decision

Predicting other annotation information when similar sequences are found

spIP00561IAK1H_ECOLI BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 1114 0.0
spIP27725IAK1H_SERMA BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 1074 0.0
spIP49079IAKH1_MAIZE BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 1023 0.0
spIP49080IAKH2_MAIZE BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 993 0.0
spIP44505IAKH_HAEIN BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D... 981 0.0
spIP37142IAKH_DAUCA BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D... 980 0.0
spIP57290IAKH_BUCAI BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE D... 974 0.0
spIP00562IAK2H_ECOLI BIFUNCTIONAL ASPARTOKINASE/HOMOSERINE ... 818 0.0

SwissProt Database

Organism Databases

LOCUS AK1H_ECOLI 820 aa linear BCT 16-OCT-2001

COLICARD — **AK1H_ECOLI**

Last_Update	March 26, 2002
Name	Bifunctional aspartokinase-Homoserine dehydrogenase I
Alternate_Names	1) Homoserine dehydrogenase 2) L-homoserine: NAD (p) + oxidoreductase
General_Function	Amino acid biosynthesis and metabolism
Specific_Function	Aspartate metabolism; Glycine metabolism; Serine metabolism; Threonine biosynthesis
Gene_Ontology	Function: # No Data; Cell Process: # No Data; Cell Component: # No Data
Pfam_Domain/Function	PF01842 ACT domain 316-384; PF00742 Homoserine dehydrogenase; PF00696 Amino acid kinase family
Similarity	In the N -terminal section; belongs to the aspartokinase family. In the C -terminal section; belongs to the homoserine dehydrogenase family
Cell_Location	Cytoplasm
Gene_Name	thrA
BlattNer_Number	b0002
Gene_Position	337-2799 (Clockwise)
Centisome_Position	0.007 minutes
Metabolic_Importance	Non_Essential
Preceding_Gene	thrL (thr operon leader peptide)
Following_Gene	thrB (homoserine kinase)
Paralogues	b3940 (metL);
Protein_Copy_Number	Unknown
# RNA_Copy_Number	Log Phase: 0.04 Stationary Phase: 0.1

EC Number Inferencer

EC Number 1.1.1.3

a set of prediction algorithms, one for each field - some simple and some as complex as general function prediction

Acknowledgements

Thanks to the 2001 summer students who worked on Proteome Analyst:
Cynthia Luk
Samer Nassar
Kevin McKee
Shannon Brooks

Thanks to some of the micro-biologists who tested Proteome Analyst and provided significant feedback:
Warren Gallin
Kathy Magor
Chris Upton

POWERED BY Sun microsystems
Proteome Analyst has received a generous equipment donation from Sun Microsystems.