2012

Characterization and annotation of hypothetical proteins

Domenico De Bernardo

CHARACTERIZATION AND ANNOTATION OF HYPOTHETICAL PROTEINS

by

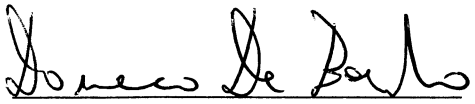
DOMENICO DE BERNARDO

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN
THE DEPARTMENT OF BIOLOGICAL SCIENCE
TO
THE SCHOOL OF GRADUATE SCIENCES
OF
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

HUNTSVILLE, ALABAMA
2012

In presenting this thesis in partial fulfillment of the requirements for a master's degree from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department or the Dean of the School of Graduate Studies. It is also understood that due recognition shall be given to me and to The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.



(Student Signature)

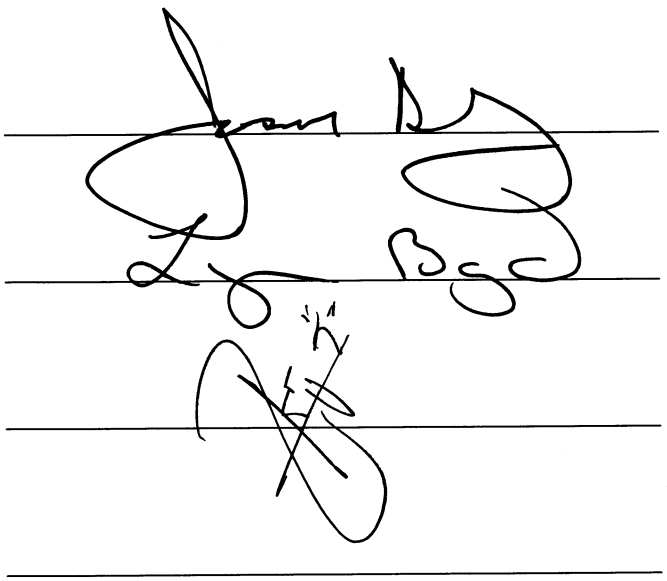
3/21/2012

(Date)

THESIS APPROVAL FORM

Submitted by Domenico De Bernardo in partial fulfillment of the requirements for the degree of Master of Science in Biological Sciences and accepted on behalf of the Faculty of the School of Graduate Studies by the thesis committee.

We, the undersigned members of the Graduate Faculty of The University of Alabama in Huntsville, certify that we have advised and/or supervised the candidate on the work described in this thesis. We further certify that we have reviewed the thesis manuscript and approve it in partial fulfillment of the requirements for the degree of Master of Science in Biological Sciences.



Committee Chair

Department Chair

College Dean

Graduate Dean

ABSTRACT
The School of Graduate Studies
The University of Alabama in Huntsville

Degree Master of Science College/Dept. Science/Biological Sciences

Name of Candidate Domenico De Bernardo

Title Characterization and Annotation of Hypothetical Proteins

The accuracy and completeness of the annotation associated to a sequenced genome is of key importance for the use of genome data. Unfortunately, the biological functions of a large amount of gene coding sequences are not determined and it is often unclear if they express proteins *in vivo*. The products of these genes are called hypothetical proteins. Experimental determination of the functions of all these proteins is unfeasible because of the time required for their characterization. The goal of this work is to develop and describe a bioinformatics tool for functional identification and annotation of hypothetical proteins. It allows parametric searches based on amino acid sequence similarities and protein structural alignments. The tool was utilized by analyzing all the hypothetical proteins identified in twelve organisms in the NCBI RefSeq database and was capable of finding thousands of possible annotations with a high level of confidence rivaling the most updated existing database.

Abstract Approval: Committee Chair

Department Chair

Graduate Dean

ACKNOWLEDGMENTS

I would like to sincerely thank Dr. Joseph Ng for his suggestion of the research topic, for his mentorship, advice and guidance throughout all the stages of the work and the completion of the thesis. In addition, I would like to thank Dr. Lynn Boyd and Dr. Luis Rogelio Cruz-Vera for participating on my thesis committee.

My biggest thank and deepest gratitude goes to my wife Adele, she always stood beside me and supported all the endeavors I have dreamed and pursued. She took care of our precious and wonderful sons Mario and Luca, while I spent most of my time between work and study.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

CHAPTER 1

INTRODUCTION

1.1 Background

The access to sequenced and assembled genomes has become an essential tool for analysis of cellular pathways, for the understanding of biological processes and for the prediction of unknown protein functions or their evolution. The accuracy and completeness of the annotations associated to a sequenced genome is the key element for the utility of the genome data. Unfortunately, even the most studied organism has a large amount of genes with a not determined function and it is often unclear whether they encode actual protein *in vivo*. The products of these genes are commonly termed hypothetical proteins, or conserved hypothetical proteins if they are homologous to other genes of unknown function (Sivashankari and Shanmughavel 2006).

As of December 15, 2011, GenBank protein database contained 146,413,798 protein sequences from ~250,000 organisms; one out of ten proteins was annotated as "hypothetical" (Benson, Karsch-Mizrachi et al. 2012).

The time necessary for the experimental characterization of these hypothetical proteins is the limiting factor for the completeness of the genome annotation. For example, *E. coli* genes with unknown function are being characterized with a rate of 20/30 per year, this means it will take many decades before the biological function of all *E-coli* proteins is established (Kolker, Makarova et al. 2004).

Bioinformatics has become a key player for the rapid and reliable prediction and determination of the molecular function of proteins and an essential tool for curator to identify and annotate proteins in databases. The most common method to infer protein function is based on amino acid sequence similarities by searching protein sequence databases with software tools such as BLAST from the National Center of Biotechnology Information (NCBI) (Bergman 2007).

Other typical bioinformatics tools to increase annotation accuracy or used when amino acid sequence similarity searches do not provide reliable results are based on phylogenetic profiles, protein-protein interactions, protein expression profiles, protein domain conservation and protein structure similarities (Sierk and Pearson 2004).

Protein function identification and annotation is a dynamic process; discoveries of cellular functions for a protein in one organism trigger an obvious update in the protein annotation for the considered organism. However, this can have a potential ripple effect on the annotation of similar proteins in different organisms. Researchers proposing an annotation for a characterized protein do not evaluate also the cascade of potential additional annotations in different organisms. Database curators are capable to reconcile the broken annotation links, but, due to the size of the databases and the high rate of publications with evidence of potential new protein functions, this can often result in a daunting task.

Hereafter, a tool, called HYPE (HYpothetical Protein Evaluation), is decribed. It was developed to analyze a large amount of hypothetical proteins, to facilitate the identification of their biological function and propose annotations. HYPE is based on NCBI BLAST for the searches of amino acid sequence similarities and NCBI VAST for

protein structural similarities. The potential benefit of using HYPE for protein annotation was evaluated against the NCBI RefSeq database, one of the most curated and accurate protein databases that is publicly available. The capability of HYPE to augment the number of annotations in RefSeq is a relevant benchmark for the utility of this tool. HYPE was exercised against all the hypothetical proteins present in twelve organisms and the results are somehow unexpected. It allowed the potential annotation of more than one hundred hypothetical proteins with a 100% positive amino acid sequence similarity match and more than a thousand with 90% positive similarities. HYPE annotation based on amino acid sequence similarity is expanded by analyzing structural similarity between proteins (Zarembinski, Hung et al. 1998). HYPE is capable of running the web based NCBI VAST tool in a recursive manner for all the hypothetical proteins and collects all the relevant data for their evaluation. The annotations are automatically proposed by HYPE and could be combined with the one found by amino acid sequence similarity searches. HYPE can be easily extend to annotate more than the analyzed twelve organisms and it is a useful tool to improve our knowledge of uncharacterized proteins, standardize their annotations; and possibly prioritize annotations based on the level of confidence in similarity matches.

1.2 HYPE high level organization

HYPE investigates potential hypothetical protein annotations based on amino acid sequence similarities and protein structural similarities. The two methods can be performed independently; the results can be combined with an ad-hoc developed post-processing program. Figure 1.1 shows a conceptual functional diagram of HYPE. The left

hand branch in Figure 1.1 shows the steps performed for the annotation of hypothetical proteins based on amino acid sequence similarities. It consists of finding the NCBI GI number (i.e. sequence identifier for all proteins in the NCBI) for all the hypothetical proteins to be analyzed, and successively of running recursively the off-line version of NCBI BLAST (i.e. blastp). BLAST searches were performed against NCBI RefSeq database, but they can be potentially run on any NCBI database. BLAST searches could be very time consuming in particular if a high number of hypothetical proteins are analyzed. However, the computational load is easily spread across multiple processors of multiple platforms for parallel computing because each hypothetical protein could be analyzed independently from one another.

After the recursive BLAST searches on all hypothetical proteins are completed, a post-process program performs data mining of BLAST results and a parametric analysis of the level of similarity between proteins. The result of the post-processing is a list of proposed annotations for the hypothetical proteins that satisfies the criteria of similarity requested. The criteria of similarities are selectable as they include the minimal percentage of amino acid sequence alignment between query and subject protein, minimal percentage of query and/or subject protein coverage (i.e. the number of amino acids aligned by BLAST with respect to the number of amino acids forming the protein), minimal amino acid length of the query protein.

The right hand branch in Figure 1.1 shows the steps performed for the annotation of hypothetical proteins based on protein structural similarity utilizing the NCBI tool VAST. Hypothetical proteins with an experimentally determined structure are found by a

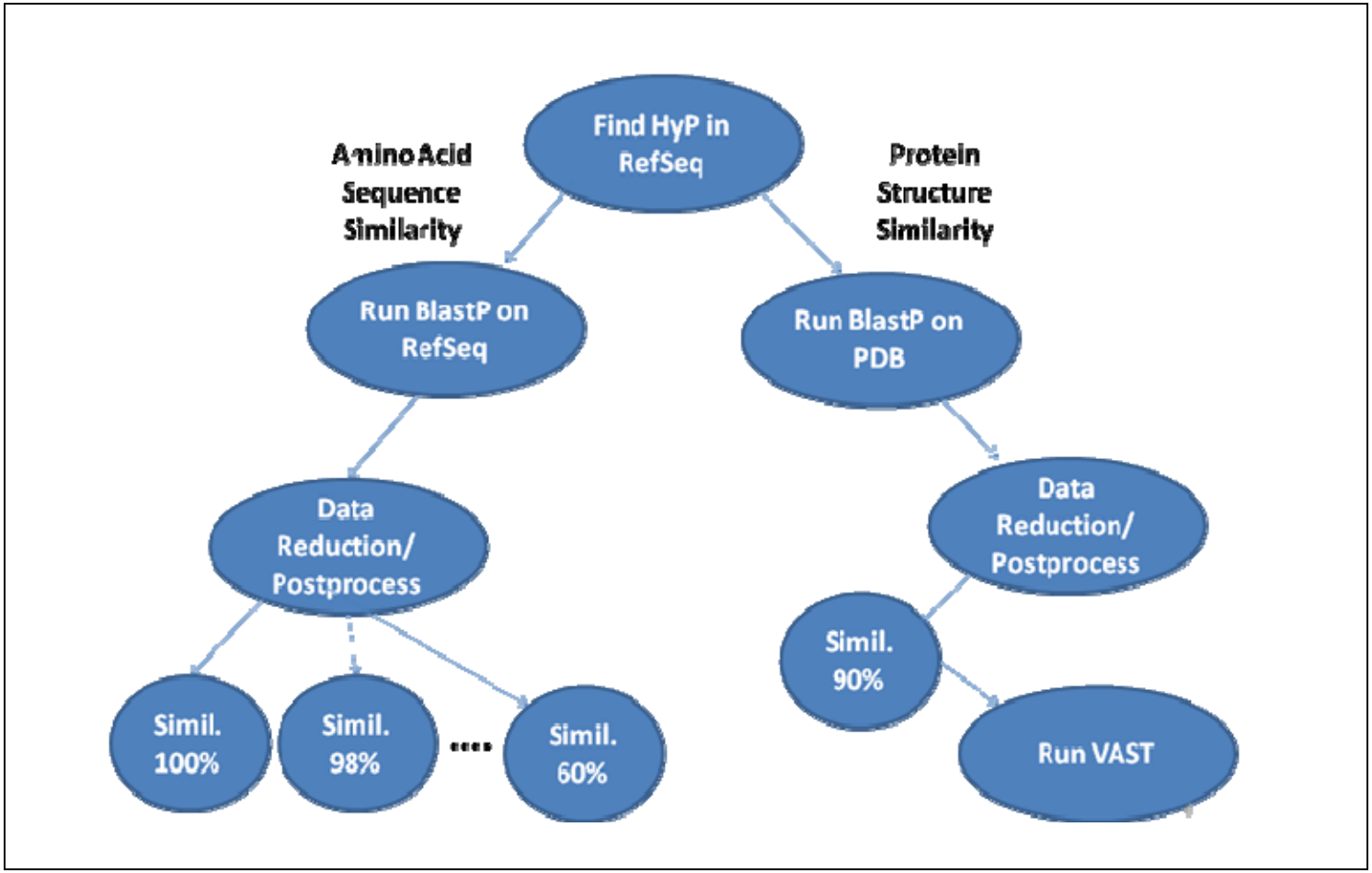


Figure 1.1. HYPE high level organization diagram. Left branch shows amino acids sequence similarity searches based on NCBI BLAST with data mining and post-processing for parametric analysis. Right branch shows protein structural alignment searches based on NCBI VAST. Protein structures are found by BLAST search on PDB database.

recursive BLAST search on protein data bank (PDB) database. It is typically rare to find hypothetical proteins with an associated structure, for this reason, the search for protein structure similarities was extended to those proteins that have an amino acid sequence alignment greater than 90% to the query protein.

For example, in Figure 1.2, query protein “hypothetical protein PF0248” in *Pyrococcus furiosus* does not have an experimentally determined structure so it cannot be analyzed with NCBI VAST. However, the BLAST run on PDB database showed that another protein from *Pyrococcus Horikoshii Ot3* (“Chain D, Ph1918 Protein”) has an associated structure and it is very similar to “hypothetical protein PF0248”. Protein structural similarity match is run on “Chain D, Ph1918 Protein”; the obtained results are likely to be applicable also to the query protein “hypothetical protein PF0248” form *Pyrococcus furiosus*. The underline assumption is that the few amino acids of difference between the two proteins have only a minimal impact on the protein 3D structure.

Because NCBI VAST is a web based tool, HYPE has an ad-hoc Perl script capable of running automatically and recursively VAST against all the hypothetical proteins without the need for the user to manually fill the protein input box in the NCBI VAST webpage. The VAST results are read from the webpage and collected in a data file. Ad-hoc post-processing script reorganizes the data and proposes the most significant annotations.

1.3 HYPE Inputs

Table 1.1 reports the twelve organisms analyzed by HYPE to assess the feasibility and the performance of the proposed annotation approach. They have sequenced and

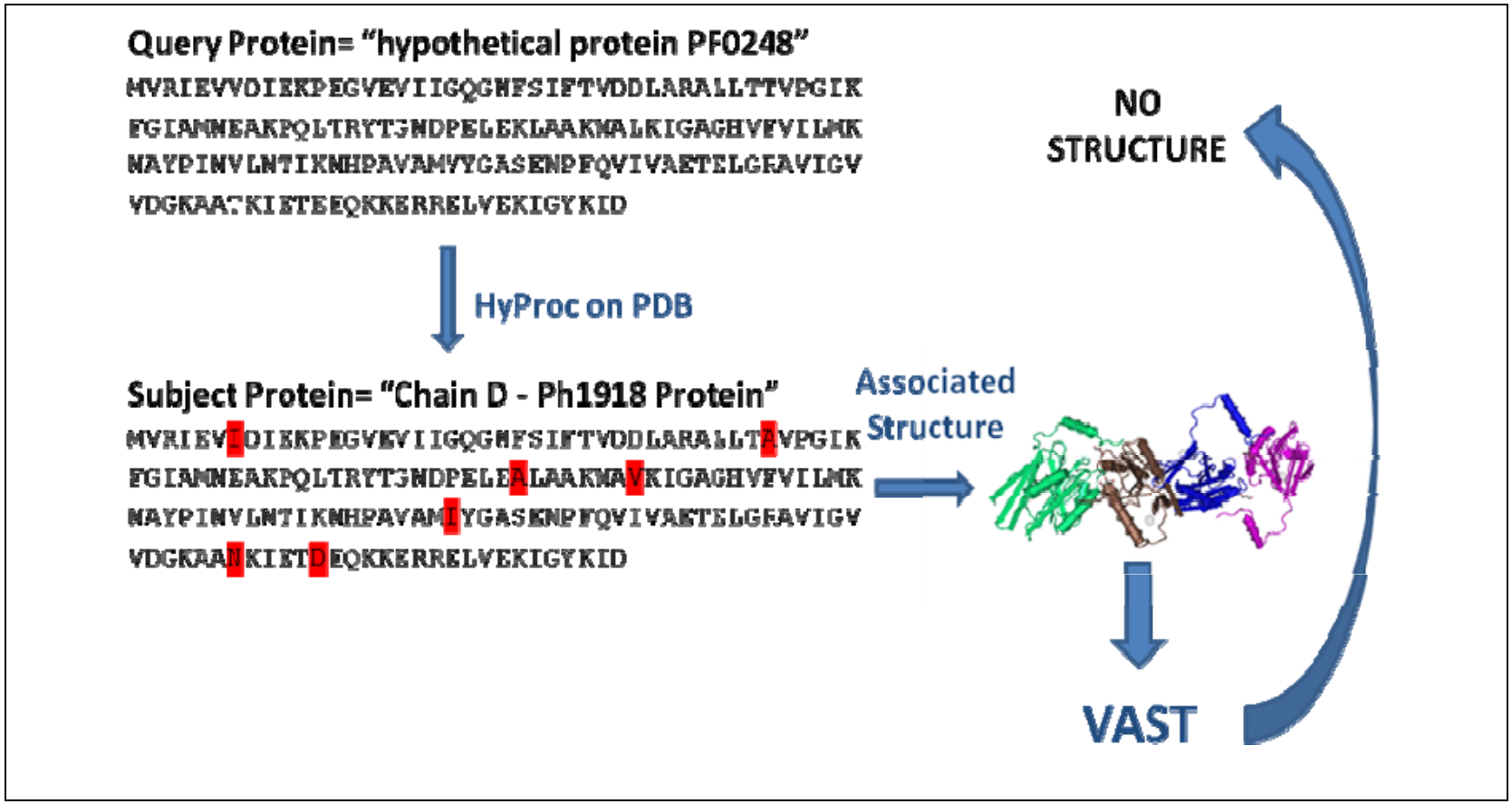


Figure 1.2. Example of PDB search for the analysis of protein structure similarities. Query protein "hypothetical protein PF0248" in *Pyrococcus furiosus* does not have an experimentally determined structure. HYPE on PDB database finds a similar protein with an associated structure, "Ph1918 Protein", in *Pyrococcus Horikoshii Ot3*. VAST search is performed on the structure of this last protein and the VAST results are linked to the initial hypothetical protein. Highlighted in red is the amino acids difference between the two proteins.

annotated genomes with size ranging from $2x10^6$ to $3x10^9$ base pairs derived from: Archea (*Thermococcus kodakarensis* (Atomi, Fukui et al. 2004), *Pyrococcus furiosus* (Robb, Maeder et al. 2001), *Methanococcoides burtonii* (Allen, Lauro et al. 2009)), Eubacteria (*Escherichia coli K12* (Blattner, Plunkett et al. 1997)), Fungi (*Saccharomyces cerevisiae* (Wood, Rutherford et al. 2001), *Magnaporthe oryzae* (Kour, Greer et al. 2011)), Plantae (*Arabidopsis thaliana* (The-Arabidopsis-Genome-Initiative 2000), *Populus trichocarpa* (Brunner, Busov et al. 2004)) and Animalia (*Danio rerio* (Sanger-Institute 2001-2012), *Caenorhabditis elegans* (C.elegans-Sequencing-Consortium 1998), *Mus musculus* (Sager-Institute 2007) and *Homo sapiens* (Muzny, Scherer et al. 2006)).

These genomes have been selected because they represent five kingdoms of life and all of them belong to existing research programs. Among these genomes, there are a combined fourty thousand hypothetical proteins encoding for novel and conserved proteins .Table 1.1 also shows for each organism the number of proteins present in NCBI RefSeq database; the number of proteins annotated as “hypothetical”; and the percentage of hypothetical proteins with respect to the total number of proteins. The organism taxonomic ID is the only input accepted by HYPE. The tool automatically provides annotations for all the hypothetical proteins present in the selected organism.

1.4 RefSeq

RefSeq is an NCBI database that was chosen because of its characteristics and non-redundancy. There is only one record for each protein, mRNA and tRNA and it is constantly updated by NCBI staff members with current knowledge of sequence data. When a sequence alignment is found with a protein in RefSeq database, the

Table 1.1. Organisms whose genome were analyzed by HYPE. The taxonomy ID associated to each organism is the one reported in NCBI. The total number of proteins for each organism was counted from the output of the NCBI offline tool "blastdbcmd" run on RefSeq database. The number of hypothetical protein was computed by searching the keyword "hypothetical" in the annotation text of each protein and the percentage of hypothetical protein is simply calculated dividing the number of hypothetical proteins with the total number of proteins.

Organism	Taxonomy ID	Number of Protein in Refseq	Number of Hypothetical Protein	Percentage of Hypothetical Protein
Thermococcus kodakarensis KOD1	69014	2306	922	40
Pyrococcus furiosus DSM 3638	186497	2125	1017	47.9
Methanococcoides burtonii - DSM 6242	259564	2273	792	34.8
E coli K12 - sub MG1655	511145	4145	21	0.5
Saccharomyces cerevisiae S288c	559292	5882	995	16.9
Magnaporthe oryzae	242507	14009	13769	98.3
Arabidopsis thaliana	3702	35372	86	0.2
Populus trichocarpa	3694	40521	268	0.7
Danio rerio	7955	27241	5180	19
Caenorhabditis elegans	6239	23906	13824	57.8
Mus musculus	10090	30045	3072	10.2
Homo sapiens	9606	34864	2852	8.2
Total		**240934**	**42798**	**19**

characterization of the protein function can be estimated by HYPE with the highest confidence level.

RefSeq is a subset of NCBI “nr” database, this has the disadvantage of obtaining a reduced number of sequence alignment hits, but the quality of the results is typically higher and the execution time for each search is faster. Current execution time is in the order of weeks for two computers running in parallel for all the forty thousand hypothetical proteins. RefSeq adopts a standard naming convention such that an automatic tool such as HYPE can exploit this characteristic to minimize human intervention in the annotation process.

The accession number in RefSeq gives an indication of the quality of the information stored in the protein records. RefSeq accession format and significance are partially reported as follows (McEntyre and Ostell 2002):

- NP_123456: “N” stands for experimentally verified protein function
- XP_123456: “X” stands for annotation computed from annotation tools
- YP_123456 : Y is typically used for organisms such as virus and bacteria

The second letter of the RefSeq accession number has the following meaning: “P” refers to the fact that the record is related to protein, “M” to mRNA, “R” to non-coding RNA transcripts including structural RNAs, transcribed pseudogenes, and others.

1.5 NCBI BLAST Overview

The sequence similarity search of HYPE is base on NCBI BLAST tool. BLAST finds regions of similarity between sequences of proteins and/or nucleotides. Its acronym stands for Basic Local Alignment Search Tool (BLAST). “Local alignment” is a key

feature of BLAST. In fact, it indicates that the comparison between two proteins is not performed along the entire length of the sequence.This allows finding regions that share high similarities and consequently similar functions in proteins that can be mostly different. This feature could be very important in the identification, for example, of conserved domains. More than one subsequence alignment can be found within a single BLAST search; each subsequence alignment is reported and can be evaluated independently from one another. On the contrary, “global alignment” refers to comparison performed along the entire length of a protein. As degree of sequence similarity declines, global alignment methods tend to miss important biological relationship (Polyanovsky, Roytberg et al. 2011).

Every local or global sequence alignment tool provides a score value to summarize in one number the quality of sequence superimposition. The score can be computed in different ways. One easy way, for example, could be to just count how many identical amino acids are aligned. BLAST uses a more complicated approach based on the fact that amino acid mismatches do not always have the same biological impact. For example, exchanging a leucine with an isoleucine does not have a big impact on either protein structure or protein chemical characteristics. This kind of substitutions is called positive substitution. However, replacing a leucine with negatively charged amino acid such as aspartic acid could potentially have high consequences on the function of the protein. These substitutions are called “negative substitution”. The matrix that maps all amino acid substitutions with an associated score are called scoring matrices or substitution matrices. BLAST can use multiple scoring matrices, the default one is called “BLOSUM62” (Henikoff and Henikoff 1992) and is represented in the next Figure 1.3.

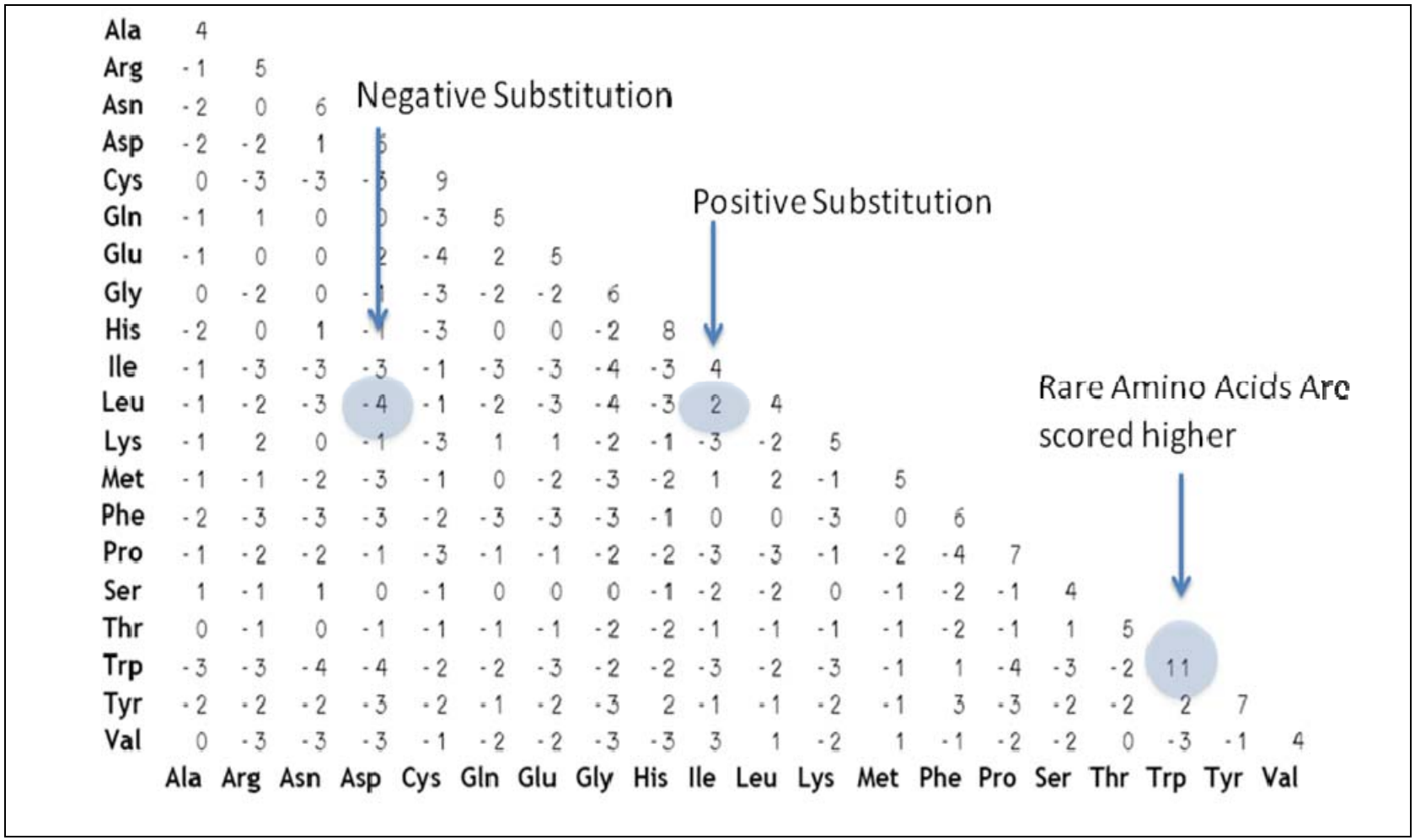


Figure 1.3. BLAST default scoring matrix Blosum62. Scoring for positive and negative substitutions is highlighted in light blue. Matches of rare amino acids have the highest scores.

Scoring coefficients have high values if the substitutions of amino acids conserve size, charge, hydrophobicity and structure of a protein. Another factor for the generation of the scoring coefficients is the “frequency” of amino acids. Frequency represents how often a particular amino acid is present in proteins. Rare amino acids are scored differently than common amino acids.

Gaps in the sequence alignment are also scored; in fact, they represent biological events such as amino acid deletion or insertion in the protein. They cannot be scored as amino acid mismatch; however they are taken into account in the scoring by a penalty.

In BLAST, gaps penalty are computed with the following formula:

Gap Penalty = G +Ln

Where G represents a constant penalty coefficient for the opening of a gap (in BLAST the value of G is 11 for proteins) (Altschul, Madden et al. 1997), “n” is the extension of the gap in terms of number of amino acids, and L is the coefficient for the penalty of gap extension (in BLAST the value of L is 1 for proteins).

1.5.1 NCBI BLAST Result Overview

The BLAST results of a particular amino acid sequence similarity match are represented in the next three figures. Each of the colored bars in Figure 1.4 represents a particular BLAST hit ordered by descending score. The color of each bar represents the value of the score divided in classes. It is possible to note that there are some proteins containing multiple alignments. The thinner gray line between two colored bars indicates

the part of the protein that was not aligned. It is evident the capability of BLAST to evaluate local region of alignment within a single protein.

The same results are also reported in the output hit list showed in Figure 1.5. The figure shows the proteins that BLAST algorithm deemed similar to the query protein by extending the alignment out as far as it can go. The list reports the following information for each BLAST hit.

- **NCBI accession number**: a unique identifier given to a sequence when it is submitted to one of the DNA repositories (GenBank, EMBL, DDBJ). The initial deposition of a sequence record is referred to as version 1. If the sequence is updated, the version number is incremented, but the accession number will remain constant (Cerami 2005).
- **Description**: information relative to the sequence indicating its function or phenotype. In case of hits in RefSeq database, the description contains, in square brackets, the name of the organism from which the sequence was derived.
- **Max Score**: Value that indicates the quality of the alignment; it is calculated from a formula that takes into account the substitution matrix (the default is BLOSUM62), the number of gaps and the length in terms of amino acids of the of the alignment.
- **Total Score**: Value computed in the same way as “Max Score” but calculated along the full protein alignment. BLAST can computes multiple local alignments within a single protein; the “Total Score” is the sum of the single alignment scores within a protein.

- **Query Coverage**: The percentage of query protein that was aligned to a protein in the database with respect to the total query protein length in term of number of amino acids.
- **E- Value**: The probability that the alignment is the results of chance alone. In other world, an E-Value of 0.01 means that there is one percent of probability that the BLAST match is due to fortunate random coincidence (Prakash and Tompa 2005).
- **Links**: A hypertext link to external resources that provide more information about the subject protein. Links can be to gene information, structures, chromosome map viewer, and public available assay.

Detail alignment results are available for all the BLAST hits shown in the output hit list of Figure 1.5. An example of detailed alignment result is showed in the next Figure 1.6. The values collected by HYPE are highlighted in light blue. These values are collected for all the BLAST hits generated by the forty thousand hypothetical proteins analyzed. In BLAST, the protein that is compared to the query protein is called subject protein. Figure 1.6 shows the following information:

- **Accession Number**: The accession number is defined previously in this paragraph.
- **Protein Name**: Description field already present in the output hit list of Figure 1.5. The description in this case is not limited in terms of number of characters. RefSeq proteins have a standard description that includes the

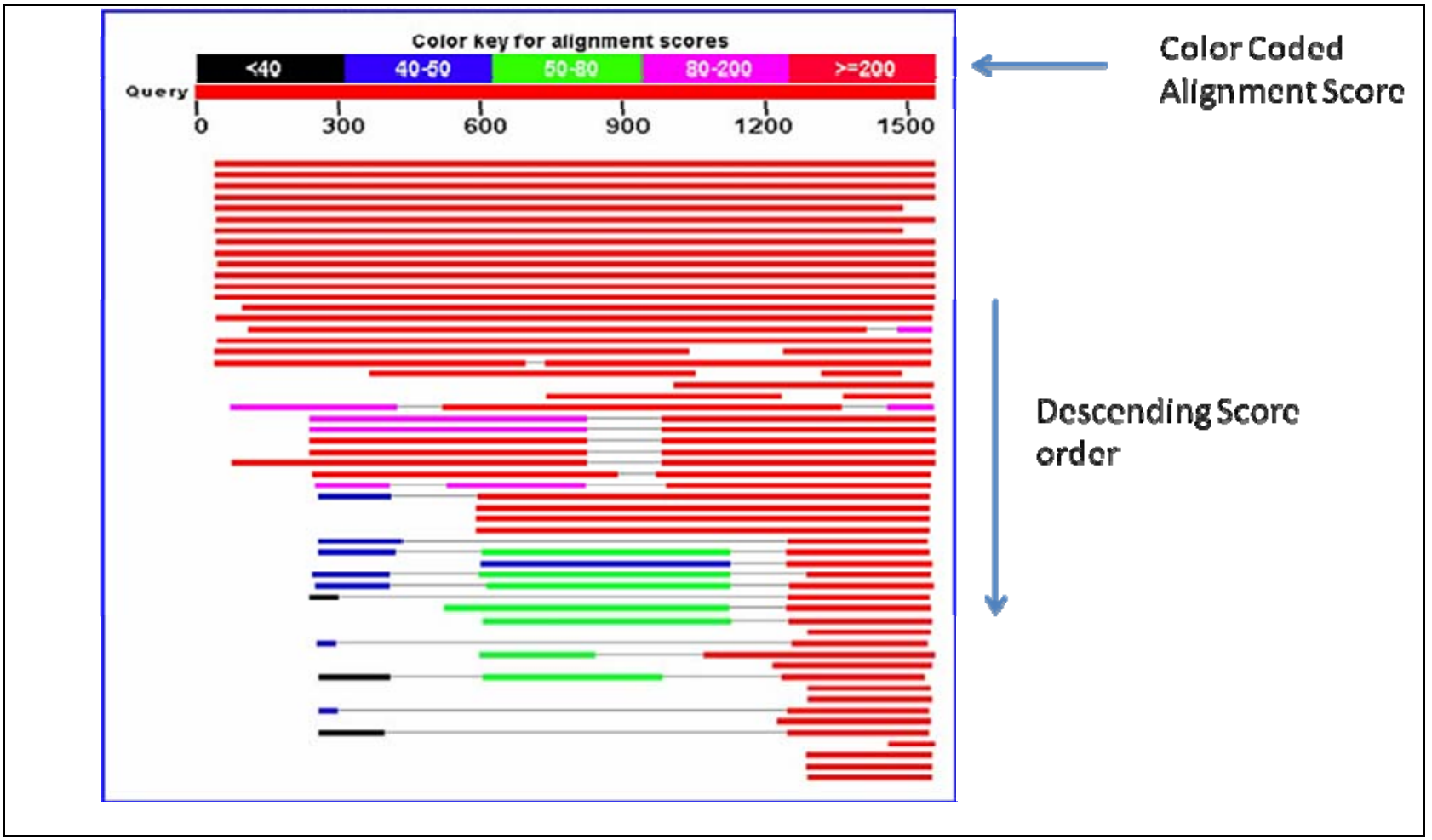
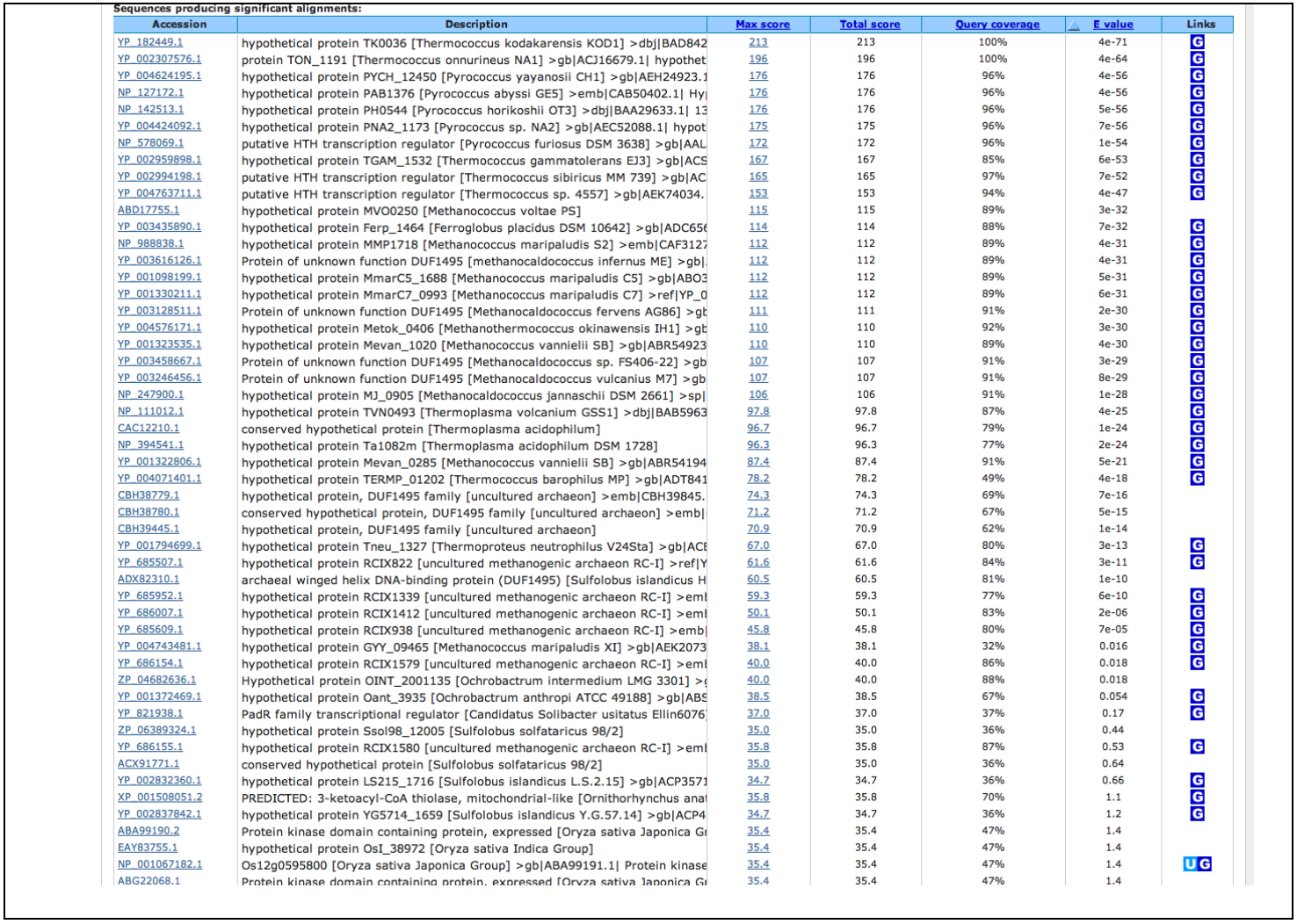


Figure 1.4. Graphical representation of BLAST alignment results. Order of results is by descending score. The BLAST hits are color coded to provide a visual representation of the significance of the results.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
YP_182449.1	hypothetical protein TK0036 [Thermococcus kodakarensis KOD1] >dbj|BAD842	213	213	100%	4e-71	G
YP_002307576.1	protein TON_1191 [Thermococcus onnurineus NA1] >gb|ACJ16679.1| hypothet	196	196	100%	4e-64	G
YP_004624195.1	hypothetical protein PYCH_12450 [Pyrococcus yayanosii CH1] >gb|AEH24923.1	176	176	96%	4e-56	G
NP_127172.1	hypothetical protein PAB1376 [Pyrococcus abyssi GE5] >emb|CAB50402.1| Hy	176	176	96%	4e-56	G
NP_142513.1	hypothetical protein PH0544 [Pyrococcus horikoshii OT3] >dbj|BAA29633.1| 13	176	176	96%	5e-56	G
YP_004424092.1	hypothetical protein PNA2_1173 [Pyrococcus sp. NA2] >gb|AEC52088.1| hypot	175	175	96%	7e-56	G
NP_578069.1	putative HTH transcription regulator [Pyrococcus furiosus DSM 3638] >gb|AAL	172	172	96%	1e-54	G
YP_002959898.1	hypothetical protein TGAM_1532 [Thermococcus gammatolerans EJ3] >gb|ACS	167	167	85%	6e-53	G
YP_002994198.1	putative HTH transcription regulator [Thermococcus sibiricus MM 739] >gb|AC	165	165	97%	7e-52	G
YP_004763711.1	putative HTH transcription regulator [Thermococcus sp. 4557] >gb|AEK74034.	153	153	94%	4e-47	G
ABD17755.1	hypothetical protein MVO0250 [Methanococcus voltae PS]	115	115	89%	3e-32	
YP_003435890.1	hypothetical protein Ferp_1464 [Ferroglobus placidus DSM 10642] >gb|ADC656	114	114	88%	7e-32	G
NP_988838.1	hypothetical protein MMP1718 [Methanococcus maripaludis S2] >emb|CAF3127	112	112	89%	4e-31	G
YP_003616126.1	Protein of unknown function DUF1495 [methanocaldococcus infernus ME] >gb|	112	112	89%	4e-31	G
YP_001098199.1	hypothetical protein MmarC5_1688 [Methanococcus maripaludis C5] >gb|ABO3	112	112	89%	5e-31	G
YP_001330211.1	hypothetical protein MmarC7_0993 [Methanococcus maripaludis C7] >ref|YP_0	112	112	89%	6e-31	G
YP_003128511.1	Protein of unknown function DUF1495 [Methanocaldococcus fervens AG86] >gb	111	111	91%	2e-30	G
YP_004576171.1	hypothetical protein Metok_0406 [Methanothermococcus okinawensis IH1] >gb	110	110	92%	3e-30	G
YP_001323535.1	hypothetical protein Mevan_1020 [Methanococcus vannielii SB] >gb|ABR54923	110	110	89%	4e-30	G
YP_003458667.1	Protein of unknown function DUF1495 [Methanocaldococcus sp. FS406-22] >gb	107	107	91%	3e-29	G
YP_003246456.1	Protein of unknown function DUF1495 [Methanocaldococcus vulcanius M7] >gb	107	107	91%	8e-29	G
NP_247900.1	hypothetical protein MJ_0905 [Methanocaldococcus jannaschii DSM 2661] >sp|	106	106	91%	1e-28	G
NP_111012.1	hypothetical protein TVN0493 [Thermoplasma volcanium GSS1] >dbj|BAB5963	97.8	97.8	87%	4e-25	G
CAC12210.1	conserved hypothetical protein [Thermoplasma acidophilum]	96.7	96.7	79%	1e-24	G
NP_394541.1	hypothetical protein Ta1082m [Thermoplasma acidophilum DSM 1728]	96.3	96.3	77%	2e-24	G
YP_001322806.1	hypothetical protein Mevan_0285 [Methanococcus vannielii SB] >gb|ABR54194	87.4	87.4	91%	5e-21	G
YP_004071401.1	hypothetical protein TERMP_01202 [Thermococcus barophilus MP] >gb|ADT841	78.2	78.2	49%	4e-18	G
CBH38779.1	hypothetical protein, DUF1495 family [uncultured archaeon] >emb|CBH39845.	74.3	74.3	69%	7e-16	
CBH38780.1	conserved hypothetical protein, DUF1495 family [uncultured archaeon] >emb|	71.2	71.2	67%	5e-15	
CBH39445.1	hypothetical protein, DUF1495 family [uncultured archaeon]	70.9	70.9	62%	1e-14	
YP_001794699.1	hypothetical protein Tneu_1327 [Thermoproteus neutrophilus V24Sta] >gb|ACE	67.0	67.0	80%	3e-13	G
YP_685507.1	hypothetical protein RCIX822 [uncultured methanogenic archaeon RC-I] >ref|Y	61.6	61.6	84%	3e-11	G
ADX82310.1	archaeal winged helix DNA-binding protein (DUF1495) [Sulfolobus islandicus H	60.5	60.5	81%	1e-10	
YP_685952.1	hypothetical protein RCIX1339 [uncultured methanogenic archaeon RC-I] >emb	59.3	59.3	77%	6e-10	G
YP_686007.1	hypothetical protein RCIX1412 [uncultured methanogenic archaeon RC-I] >emb	50.1	50.1	83%	2e-06	G
YP_685609.1	hypothetical protein RCIX938 [uncultured methanogenic archaeon RC-I] >emb|	45.8	45.8	80%	7e-05	G
YP_004743481.1	hypothetical protein GYY_09465 [Methanococcus maripaludis XI] >gb|AEK2073	38.1	38.1	32%	0.016	G
YP_686154.1	hypothetical protein RCIX1579 [uncultured methanogenic archaeon RC-I] >emb	40.0	40.0	86%	0.018	G
ZP_04682636.1	Hypothetical protein OINT_2001135 [Ochrobactrum intermedium LMG 3301] >g	40.0	40.0	88%	0.018	
YP_001372469.1	hypothetical protein Oant_3935 [Ochrobactrum anthropi ATCC 49188] >gb|ABS	38.5	38.5	67%	0.054	G
YP_821938.1	PadR family transcriptional regulator [Candidatus Solibacter usitatus Ellin6076]	37.0	37.0	37%	0.17	G
ZP_06389324.1	hypothetical protein Ssol98_12005 [Sulfolobus solfataricus 98/2]	35.0	35.0	36%	0.44	
YP_686155.1	hypothetical protein RCIX1580 [uncultured methanogenic archaeon RC-I] >emb	35.8	35.8	87%	0.53	G
ACX91771.1	conserved hypothetical protein [Sulfolobus solfataricus 98/2]	35.0	35.0	36%	0.64	
YP_002832360.1	hypothetical protein LS215_1716 [Sulfolobus islandicus L.S.2.15] >gb|ACP3571	34.7	34.7	36%	0.66	G
XP_001508051.2	PREDICTED: 3-ketoacyl-CoA thiolase, mitochondrial-like [Ornithorhynchus anat	35.8	35.8	70%	1.1	G
YP_002837842.1	hypothetical protein YG5714_1659 [Sulfolobus islandicus Y.G.57.14] >gb|ACP4	34.7	34.7	36%	1.2	G
ABA99190.2	Protein kinase domain containing protein, expressed [Oryza sativa Japonica Gr	35.4	35.4	47%	1.4	
EAY83755.1	hypothetical protein OsI_38972 [Oryza sativa Indica Group]	35.4	35.4	47%	1.4	
NP_001067182.1	Os12g0595800 [Oryza sativa Japonica Group] >gb|ABA99191.1| Protein kinase	35.4	35.4	47%	1.4	U G
ABG22068.1	Protein kinase domain containing protein, expressed [Oryza sativa Japonica Gr	35.4	35.4	47%	1.4	

Figure 1.5. BLAST output hit list reporting proteins analyzed by BLAST. Each hit is represented by the accession number, a description and parameters representing the quality of the alignment: BLAST score, query protein coverage, e value and links to external available information.

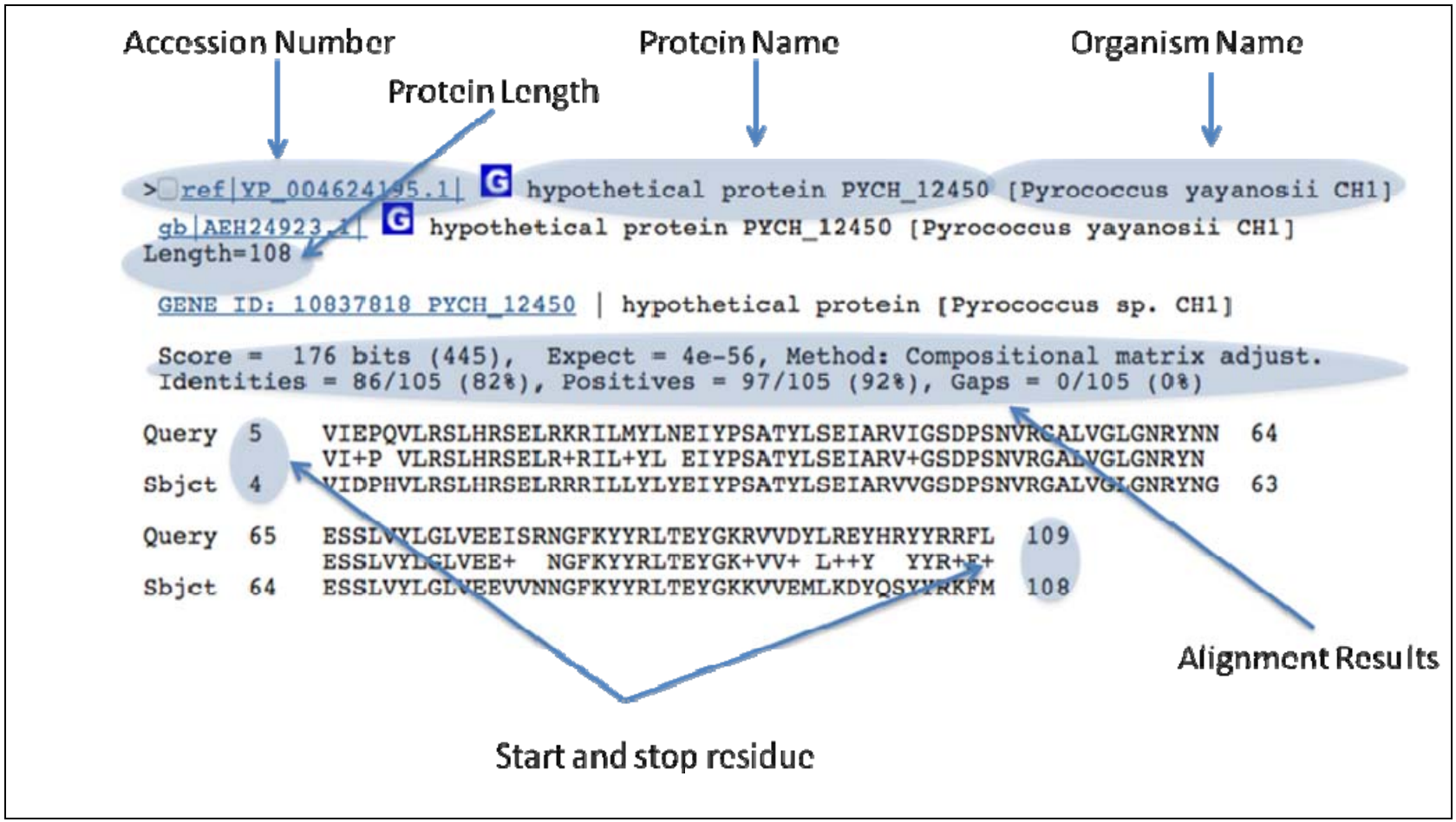


Figure 1.6. Detail alignment result of a BLAST hit. The value highlighted in light blue are collected by HYPE and organized, together with other information, in a post processing file for parametric analysis calculation.

protein name and the organism where the protein was found. The organism name is identified in square brackets. The presence of square brackets is the marker in HYPE for the collection of the name of the organism.

- **Length**: Number of amino acids forming the subject protein.
- **Start Residue**: The first amino acid in the query/subject protein from which the BLAST alignment score was computed. In particular, with reference to Figure 1.6 the query protein was aligned from the fifth amino acid. The subject protein was aligned starting from the fourth amino acid.
- **Stop Residue**: The last amino acid in the query/subject protein ending the BLAST alignment match. In particular, with reference to Figure 1.6 the query protein was aligned until amino acid 109; the subject protein was aligned till the last amino acid (i.e. 108).
- **Alignment Results**: This part contains the most important information on the quality of the alignment match performed by BLAST. The following sub-bullets describe their meaning:
    - **Score**: This is the BLAST computed score for the considered alignment. In parenthesis there is the raw score obtained by the sum of the score values for each amino acid alignment as defined in substitution matrix (BLOSUM62). The number and length of gaps is also considered in the calculation of the raw score. The BLAST score is obtained by the normalization of the raw score.
    - **Expect**: This is the e-value already described at the beginning of this paragraph

- **Method**: The computational method indicates the algorithm used for score computation.
- **Identities**: Number of amino acid aligned with identical residue. In parenthesis there is the percentage of identical amino acids with respect of total number of aligned amino acids.
- **Positives**: Number of amino acid aligned with identical residue or amino acid with positive substitution. Positive substitution refers to the fact that two amino acids can be interchanged without big impact on structure or function of the protein. In parenthesis there is the percentage of positive substitutions with respect of total number of aligned amino acids.
- **Gaps**: Number of amino acid not aligned. They usually represent biological events such as insertion of deletion. In parenthesis there is the percentage of gaps calculated by the number of gaps divided the total number of aligned amino acids.

1.6 NCBI VAST Overview

The structural similarity search of HYPE is base on NCBI VAST tool. VAST stands for Vector Alignment Search Tool (Gibrat, Madej et al. 1996), as expressed by its name, it is based on vector alignment. This means that each protein included in the PDB database is transformed in a series of vectors representing their secondary structure. Both alpha helices and beta sheets are transformed in vectors while loop structures in the protein are discarded. Each alpha helix is represented by a vector passing in the center of the helix and with the orientation given by the C-terminal and N-Terminal of the helix.

Beta sheets are handled in a similar way such that all the atomic coordinates of the atoms forming the protein are discarded reducing greatly the complexity of the structural similarity match.

Once all the proteins are transformed in set of vectors, VAST proceed to compare two proteins with a method illustrated in Figure 1.7. Protein 1 and protein 2 are formed by 4 and 5 vectors respectively. VAST tries to overlay the two proteins in all the possible ways (i.e. alignment 1, alignment 2 alignment 3 …) by also changing the numbers of vectors to be compared. The structure that superimposes with the highest accuracy receives the greatest VAST score.

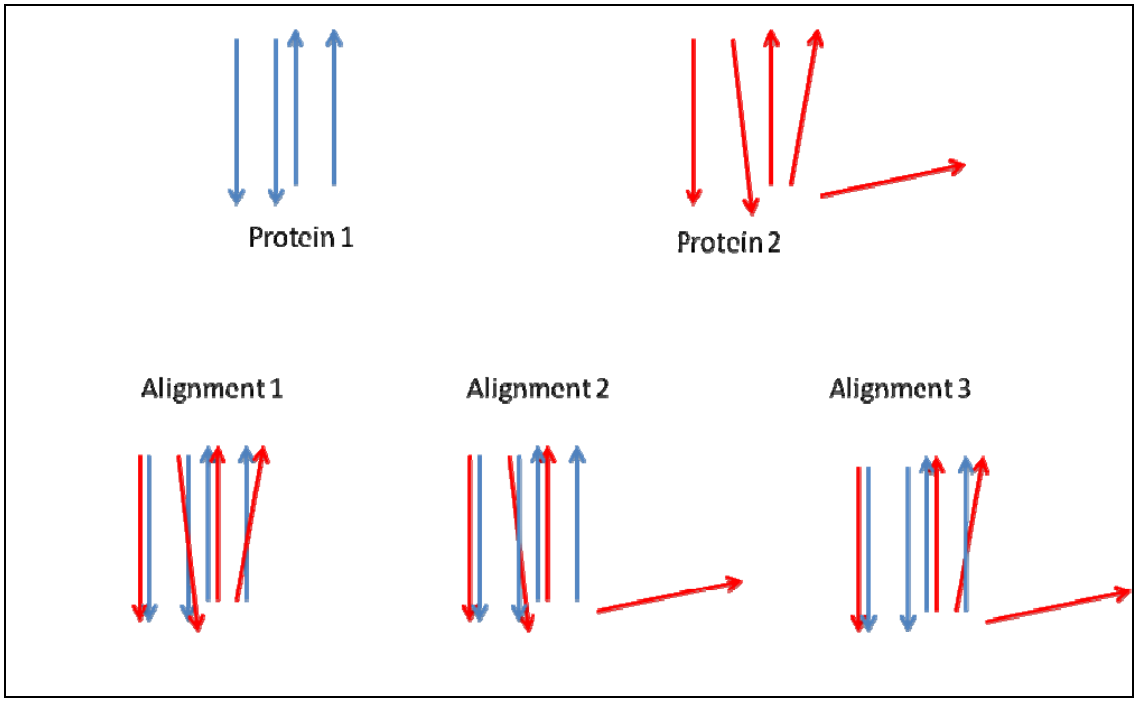


Figure 1.7. Approach for vector superposition in VAST structural similarity match. Protein 1 and protein 2 are vectorialized with vectors passing through each alpha helix and beta sheet. The vectors composing the two proteins are successivenly superimposed with different alignments configurations. The superimposition receiving the maximum VAST score will be presented by VAST as the structural alignment result.

This method is not an optimal approach for the determination of structural similarities, however regardless of its “simplicity”, it provides an accurate and fast answer to the question of structural similarity (Shapiro and Brutlag 2004).

The usefulness of structural similarity methods is important when protein similarity is not detected by traditional method such as BLAST. Sequence always specifies conformation, but conformation does not specify sequence (Anfinsen 1973). There can be structures that were conserved to a much greater extent than amino acid sequences; this conservation could give hints to the function of a protein.

1.6.1 NCBI VAST Results

VAST can report structural similarity results in different ways. Figure 1.8 shows the graphical results obtained for structure with PDB ID equal to 1XNE. The first bar shows the query protein conformation, number of amino acids and specific hits to particular domains. The subsequent bars represent other proteins, with the PDB identifier reported in blue on the right side, that are structurally similar to the query protein according to the VAST algorithms. The discontinuity in the bar represents region of the protein that do not overlay with the query protein, this gives a visual indication if the alignment is global or limited to some part of the protein or a particular domain.

It is also interesting to observe visually the overlay of one protein to another. This can be performed by checking the check box close to the protein of interest and push the “View 3D alignment” button. The protein viewer Cn3D (Madej, Addess et al. 2012) is launched and the two proteins are showed as represented in Figure 1.9. The figure illustrates the visualization of the overlap between the structures of protein 1B54 with

protein 1CT5 and 3CPG. In the first case, the two proteins share a great similarity in terms of amino acid sequence and it is not a surprise that they overlap very well (red tubes). In the second case, the proteins 1B54 and 3CPG have much less number amino acids in common (blue tubes), anyhow the structure is very similar. This is the case where tool such as VAST, could provide great insight to the function of a protein that cannot be highlighted with classical sequence similarity searches.

1.6.2 VAST Data Information

The graphical information presented in Figure 1.8 can be rendered as a table of values indicating the major statistical parameter of the structural alignment. The VAST results are described below as reported from the VAST webpage with reference to Figure 1.10.

- **PDB**: Protein data bank identifier of the structure.
- **C**: The name of the PDB chain In case of protein 2Z0T in figure, the chain name is A.
- **D**: Identifier for the Molecular Modeling Database (MMDB), a database in NCBI that contains experimentally determined structures obtained from the PDB. Protein 2Z0T, in figure, does not have an identifier in MMDB.
- **Ali. Len**: structural alignment length that is equivalent to the number of equivalent pairs of C-alpha atoms superimposed between the two structures.
- **SCORE**: The VAST structure-similarity score. This value is related to the number of secondary structure elements superimposed and the quality of that

Figure 1.8. Graphical representation of structural similarity matches in NCBI VAST.

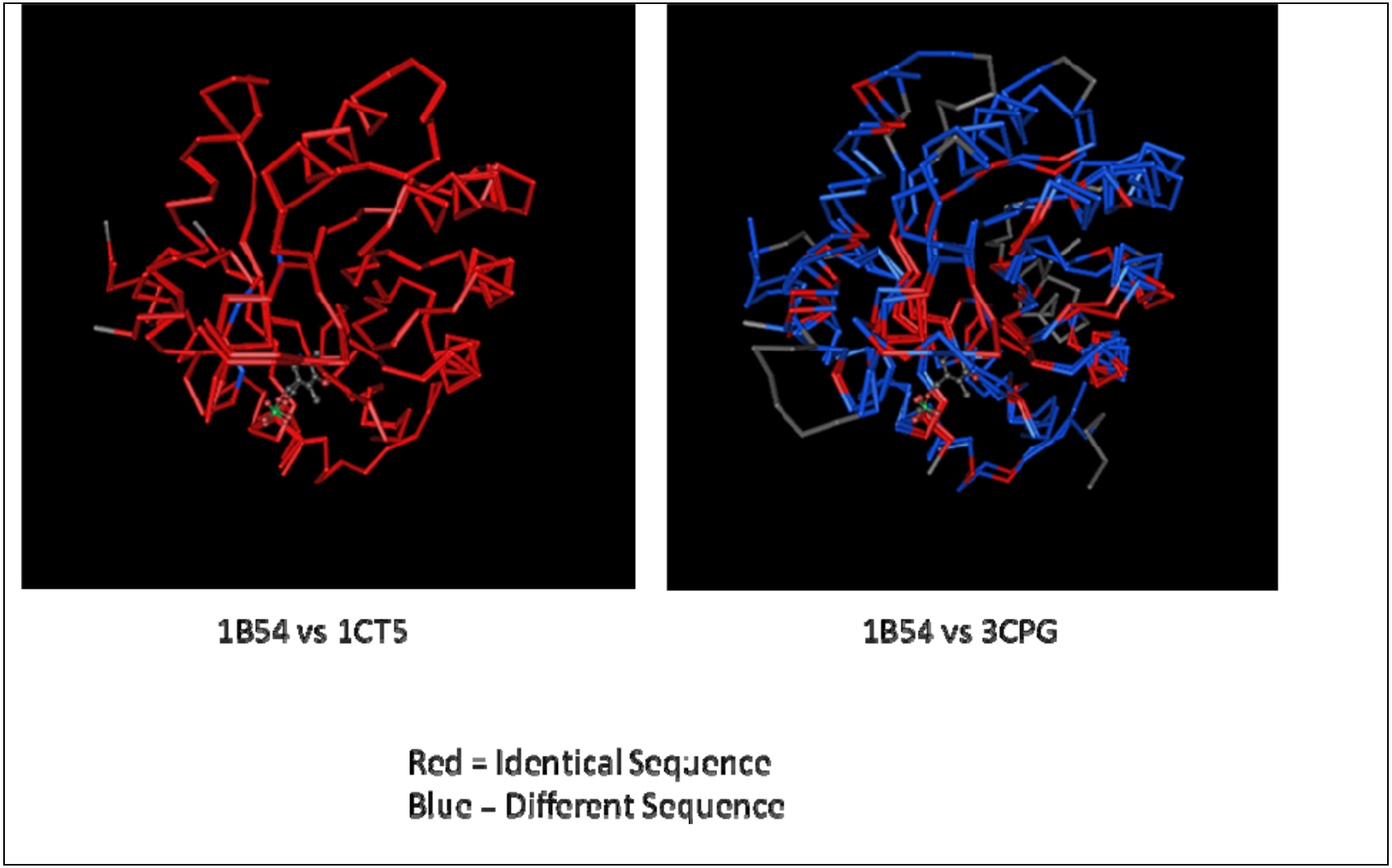


Figure 1.9. Visualization of the overlap between the structures of protein 1B54 with protein 1CT5 and 3CPG.

superposition. Higher VAST scores correlate with higher structural similarity (Madej, Gibrat et al. 1995).

- **E-VAL:** The probability that the two structures are aligned by chance only. If the e-value is 0.01, there one possibility over one hundred that the structural alignment is due by pure chance (Wrabl and Grishin 2008).
- **RMSD**: The root mean square superposition residual in Angstroms. This number is calculated after optimal superposition of two structures, as the square root of the mean square distances between equivalent C-alpha atoms. The RMSD value scales with the extent of the structural alignments and that this size must be taken into consideration when using RMSD as a descriptor of overall structural similarity.
- **%Id**: Percentage of identical residues in the aligned sequence region. This is a measure of amino acid sequence similarity in the parts of the proteins that have been superimposed.
- **MMDB date**: the date when the structure was available in the MMDB database.
- **LHM:** Loop Hausdorff Metric. A Loop Similarity measure that shows how well two structures conform to each other in the loop regions, after structural superposition. The "loop regions" the parts of the structures between aligned secondary structure elements (helices and strands) are not aligned by VAST algorithms. LHM is measured in Angstroms, with a smaller value indicative of greater similarity. (Panchenko and Madej 2004).

- **GSP:** Gapped Score. A combination (algebraic) score that uses RMSD, aligned length, and the number of gapped regions in the alignment. A smaller gapped score correlates with greater similarity (Kolodny, Koehl et al. 2005).

  Description: A text describing the annotation in the PDB database.

Total related structures: 618; 1 - 60 of 98 representatives from the Medium redundancy subset displayed. Page: 1

Click to: Check All Uncheck All

PDB C D	Ali. Len	Score	E_Val	Rmsd	%Id	MMDB Date	LHM	GSP	Description
2Z0T A	93	13.2	10e-11.3	1.7	32.3	11/2007	2.1	2.0	Crystal Structure Of Hypothetical Protein Ph0355
3S9X A	80	10.6	10e-6.7	2.1	15.0	08/2011	NA	NA	High Resolution Crystal Structure Of Asch Domain From Lactobacillus Crispatus Jv V101
2DP9 A	68	8.0	0.0062	2.1	16.2	12/2006	5.7	3.4	Crystal Structure Of Conserved Hypothetical Protein Ttha0113 From Thermus Thermophilus Hb8
1T62 B	67	10.7	10e-6.5	1.4	16.4	07/2004	5.5	2.3	Crystal Structure Of Conserved Hypothetical Protein [gi:29377587] From Enterococcus Faecalis V583
2GBS A	60	8.1	0.0051	1.8	8.3	05/2006	7.9	3.4	Nmr Structure Of Rpa0253 From Rhodopseudomonas Palustris. Northeast Structural Genomics Consortium Target Rpr3
1TE7 A	59	8.1	0.0019	2.2	20.3	01/2005	4.4	4.2	Nmr Solution Structure Of The 14kda Hypothetical Protein Yqfb From Escherichia Coli
1ZCE A	58	7.6	0.0454	2.0	17.2	05/2005	7.4	3.8	Crystal Structure Of The Hypothetical Protein Atu2648 From Agrobacterium Tumefaciens, Northeast Structural Genomics Target Atr33
3IUW A	57	8.4	10e-4.5	1.1	22.8	09/2009	6.2	2.1	Crystal Structure Of Activating Signal Cointegrator (Np_814290.1) From Enterococcus Faecalis V583 At 1.58 A Resolution
2P5D A	52	7.9	0.0082	1.4	23.1	10/2007	7.0	3.0	Crystal Structure Of Mjecl36 From Methanocaldococcus

Figure 1.10. Tabular representation of structural similarity matches in NCBI VAST.

CHAPTER 2

HYPE SEQUENCE SIMILARITY

2.1 Materials and Methods

2.1.1 Protein Databases Generation

HYPE was evaluated against two databases. The first is the whole NCBI RefSeq database and the second was obtained from RefSeq by removing all the organisms not present in Table 1.1. RefSeq was simply downloaded from NCBI ftp server: ftp://ftp.ncbi.nih.gov/refseq/release/. The tool was exercised on Release 50 of the database dated November 8, 2011 containing more than 16,000 organisms and more than 13,000,000 proteins (Pruitt, Tatusova et al. 2012). The database containing only the selected organisms of Table 1.1 was generated with the following steps:

1) Search the NCBI “Entrez” protein database (Sayers, Barrett et al. 2012) with the taxonomic ID of the organism to be added to the database. For example, to search all proteins in *Thermococcus kodakarensis KOD1* the input box must be filled with: “txid69014 [ORGN]", where 69014 is the taxonomic ID as reported in Table 1.1. This search returns, for the selected organism, all proteins recorded in NCBI and the total number of proteins present in RefSeq.

2) Create a file containing all NCBI GI number (i.e. sequence identifier for all proteins in the NCBI) of all proteins belonging to the selected organism. This is achieved by selecting the tag “Send To” in the NCBI result webpage of the previous step and

choosing the format: “GI List”. The generated file is not limited to proteins contained in RefSeq, but includes all the protein stored in all NCBI databases.

3) Create an aliased blast database using the NCBI program “blastdb_aliastool” (Bethesda 2008) with input the file generated in step 2. The execution command line is: “blastdb_aliastool -gilist GIFile -db RefSeq -out Subset_RefSeq -title Subset_RefSeq”. The file “GIFile” is created in step 2 and “Subset_RefSeq” is the chosen name of the newly created database. Even if the list of all proteins NCBI GI obtained in step 2 is not limited to proteins contained in RefSeq, the result of this third step includes only RefSeq proteins because the blastdb_aliastool command was run on RefSeq database by utilizing the parameter: “-db RefSeq”.

2.1.2 Search for Hypothetical Proteins

The search of all the hypothetical proteins for a given organism is performed by the use of a program written in “Perl” (Schwartz and Christiansen 1997) and is based on the NCBI tool “blastdbcmd“ (Bethesda 2008). Part of the source code is reported in Figure 2.1. The program is divided in two parts. The first part extracts the name of all the proteins of an organism with a given taxanomic ID. The second part selects the NCBI GI number of all the hypothetical proteins based by the presence in the protein name of the word “Hypothetical“. The hypothetical proteins found by HYPE are recorded into a file containing the following information: NCBI GI number, number of amino acids in the protein, protein name.

```
######### MAIN PROGRAM
foreach $index (@taxid2run) {
	my $taxid=$taxid_array[$index];
	my $name_spec=$name_array[$index];
	my $type_spec=$type_array[$index];
	print "Running $taxid $name_spec $type_spec\n" ;

	my $allgifilename = $taxid.'_'.$database.'_allgi.txt';
	my $hypgifilename = $taxid.'_'.$database.'_hypgi.txt';

	############## Extraction from database of all NCBI GI of organism with given taxid
	my $counter_all=0;
	if ($generate_gi==1){
		my $system_string = 'blastdbcmd -db ./db/'.$database.' -entry all -outfmt "%T
%g %l %t" | awk \'{ if ($1 == '.$taxid.') {print $0 } } \' > '.$allgifilename;
		my $status = system($system_string);
	}
	else{
		print "no gi generation\n"
	}

	############### Select only hypothetical proteins
	open (infile_all, $allgifilename) or die "Cannot open $allgifilename\n";
	open (outfile_hyp, ">$hypgifilename");

	my $counter_hyp=0;
	while (<infile_all>) {
		my $line =$_;

		if ($line =~ /hypothetical/) {
			$line =~ s/$taxid//;
			$counter_hyp = $counter_hyp+1;
			print outfile_hyp "$counter_hyp  $line";
		}
	}
	close(infile_all);
	close(outfile_hyp);
}
```

Figure 2.1. Extract of the source code for the generation of the database containing a subset of the organisms present in NCBI RefSeq.

2.1.3 Protein Sequence Similarity Search

The search for protein similarity match is performed with the offline version of NCBI BLAST, called “blastp”. All the hypothetical proteins selected in the previous steps for a given organism are searched against the whole RefSeq database and against the subset of RefSeq database containing only the organisms of Table 1.1. A Perl script, partially reported in Figure 2.2, was written to reformat the file generated in the previous step to the layout required by NCBI blastp**.**

BLAST searches were limited to an “e value” (Prakash and Tompa 2005) less than 0.0001. The output of the blastp command is in the standard web-based BLAST output format

2.1.4 Data Mining

The large amount of data generated by running blastp with the default parameters needs to be processed to retrieve only the information necessary for the functional identification of the hypothetical proteins and consequently for their annotations. For example, the details showing the alignment along the protein sequence are not necessary to be retrieved; on the contrary the statistical characteristic of the alignment must be collected. This data mining process is performed by a Perl script that reads the blastp standard output and produces a “csv” file (comma separated variable) containing in each row the sequence alignment information that have an e-value less than 0.0001 for each hypothetical protein of an organism with a specific taxonomic ID.

```
######### MAIN PROGRAM
foreach $index (@taxid2run) {
	############## GENERATION OF AllExtractedHyp.txt file
	open (outfile_hyp, ">AllExtractedHyp.txt");
	$taxid=$taxid_array[$index];
	my $name_spec=$name_array[$index];
	my $type_spec=$type_array[$index];
#	print "Extracting $taxid $name_spec $type_spec\n" ;

	my $hypgifilename = $taxid.'_'.$database.'_hypgi.txt';


	############### Select only hypothetical protein
	open (infile_all, $hypgifilename) or die "Cannot open $hypgifilename\n";

	my $counter_gi=0;
	while (<infile_all>) {
		my $line =$_;
		$line =~ /(\d+)\s+(\d+)\s+/;
		print outfile_hyp "$2\n";
		$counter_gi=$counter_gi+1;
	}
	print "Number of hyp gi=$counter_gi in $taxid\n";
	close(infile_all);
	close(outfile_hyp);
	############## END GENERATION OF AllExtractedHyp.txt file

	############## Run blastp on the newly generated AllExtractedHyp.txt file
	if ($run_blast==1){

		my $blastresult_filename = $taxid.'_'.$database.'_blast_results.txt';
		my $system_string = 'blastp -query AllExtractedHyp.txt -db ./db/'.$database.'
-evalue 0.0001 >'.$blastresult_filename;

		print "Running:$system_string\n";
		my $status = system($system_string);
	}
	else{
		print "no Blast results generation\n"
	}
}
```

Figure 2.2. Extract of the source code for the execution of the offline version of NCBI BLAST.

The data stored in the post-processed file are highlighted in the extract of the HYPE source code in Figure 2.3. The meaning of the fields is reported below.

- **Query GI**: NCBI unique identifier for the query protein.
- **Query Ref**: RefSeq accession number for the query protein.
- **Query Name**: The name of the query protein extracted from the RefSeq protein description field.
- **Query Type**: Organism name that contains the query protein extracted from the RefSeq protein description field. Standard RefSeq naming convention indicates the organism name in square brackets.
- **Percentage Query Coverage**: The percentage of amino acids of the query protein aligned by BLAST with respect to the total BLAST alignment length.
- **Percentage Subject Coverage**: The percentage of amino acids of the subject protein aligned by BLAST with respect to the total BLAST alignment length.
- **Percentage Similarity**: The percentage of amino acids aligned by BLAST between query and subject protein with respect to the total alignment length. This value contained also the amino acids with positive substitution already defined in the previous paragraphs (see Figure 1.3).
- **Percentage Gaps**: The percentage of alignment gaps with respect to the total BLAST alignment length.
- **Subject Ref**: RefSeq accession number for the subject protein.
- **Subject Name**: The name of the subject protein extracted from the RefSeq protein description field.

- **Subject Type**: Organism name that contains the subject protein extracted from the RefSeq protein description field. Standard RefSeq naming convention indicates the organism name in square brackets.
- **Subject Score**: Alignment score as reported by the BLAST search.
- **Subject Expected**: Probability that the alignment is due to chance only. This number is reported by BLAST.
- **Number of Subject Identities**: Number of identical amino acids present in the alignment.
- **Number of Subject Positives**: Number of identical amino acids present in the alignment plus the number of positive amino acid substitutions.
- **Number of Subject Gaps**: Number of alignment gaps in the subject protein.
- **Number of Amino Acids Aligned** (including gaps): BLAST total length of local alignment in terms of amino acids.
- **Subject Start**: Amino acid number counted from the start of the subject protein where the BLAST alignment started.
- **Subject End**: Amino acid number counted from the start of the subject protein where the BLAST alignment ended.
- **Subject Length**: Subject protein length in terms of amino acids.
- **Query Start**: Amino acid number counted from the start of the query protein where the BLAST alignment started.
- **Query End**: Amino acid number counted from the start of the query protein where the BLAST alignment ended.
- **Query Length**: Subject protein length in terms of amino acids.

- **Hypothetical Flag**: Flag indicating if the subject protein is a hypothetical protein. A hypothetical protein is recognized if there is the keyword "Hypothetical" in the name.
- **Same Type Flag**: Flag indicating that the subject protein is contained in the same organism of the query protein.

The information concerning the query protein is retrieved from the default BLAST output file by reading the keyword "Query=". For a single query protein, the BLAST output could report several alignments either within a single subject protein (multiple local alignments) or toward different subject proteins. The query protein information must be saved in a variable and used to generate the HYPE report for every proposed annotation. The information concerning the query protein ends when in the BLAST output file is present the keyword "Length=" followed by a number, indicating the length of the query protein.

Because the keyword "Length" is used also for the subject protein, an internal consistency and synchronization check of the information is performed within HYPE. Figure 2.4 shows an extract of the HYPE source code for the handling of query protein information. When a problem in the synchronization of the protein data is observed, a new output file is generated containing the error code and the name of the query protein where the error was discovered.

```
######### MAIN PROGRAM
foreach $index (@taxid2run) {

        $taxid=$taxid_array[$index];
        $taxid=$taxid_array[$index];
        my $name_spec=$name_array[$index];
        my $type_spec=$type_array[$index];
        my $resultfilename = $taxid.'_'.$database.'_blast_results.txt';
        my $outputfilename = $taxid.'_'.$database.'_blast.csv';
        #print "index=$index, taxid=$taxid inputfilenae=$resultfilename\n";

        open (infile_all, $resultfilename) or die "Cannot open     $resultfilename\n";
        open (outfile_res, "> $outputfilename");
        print outfile_res "Query Gi, Query Ref, Query Name, Query Type,Percentage Query Coverage,Percentage
                Subject Coverage, Percentage Similarity, Percentage Gaps,Subject Ref,Subject Name,Subject
                Type,Subject Score, Subject Expected,Number of Subject Identities, Number of Subject
                Positives, Number of Subject Gaps,Number of Amino Acids Aligned (including gaps),
                Subject Start,Subject End,Subject Length,Query Start,Query End,Query Length,
                Hypothetical Flag,Same Type Flag\n";
        my $counter_gi=0;
        $delta_old=0;
        $discarded_hit_counter=0;
        $counter_exit=0;

        while (<infile_all>) {
                my $querylines =$_;

                ########## Read query data
                if( $querylines =~ /Query= gi\|(\d+)\|ref\|(\w+\.\d*)\|/){
                        $query_gi=$1;
                        $query_ref=$2;
```

Figure 2.3. HYPE Data mining source code. Information contained in the output file.

```
#####################################################################
#### After reading the keyword "Query=" , continue reading lines
#### until keyword Length="
#####################################################################
    while (<infile_all>) {
            $querylines .= $_;
            #### Check end of query information (length is the last query information available)
            if( $querylines =~/Length=(\d+)/){
                    $query_length = $1;
                    last;
            }
    }
    $querylines =~ s/\n/ /g;
    if($querylines=~ /Query= gi\|\d+\|ref\|\w+\.\d*\|(.*)\[.*/){
            $query_name=$1;
            $query_name =~ s/,/-/g;
            #print "query name=$query_name\n";
    }
    else{die "ERROR: Cannot Read Query Name (1)";}
    if ($querylines =~/\[(.*)\]/){
            $query_type = $1;

            # sometimes there are two square brakets
            if ($query_type =~/\[(.*)$/){
                    $query_type = $1;
            }
    }
    else{die "ERROR: Cannot Read Query Type";}
    $counter_gi=$counter_gi+1;
```

Figure 2.4. HYPE Data mining source code. Rules for retrieving Query protein information.

38

Figure 2.5 shows the source code for the collection of data concerning the subject protein. In this case, the code must handle multiple subject proteins; the start of the information concerning the subject protein is recognized by the keyword “>ref”. The same keyword delimits also the data between two different BLAST hits. The end the information concerning the subject proteins is delimited by the keyword “Lambda K H”.

HYPE does not collect data for proteins with multiple alignments. When multiple alignments are found, HYPE reports only the protein name of the query and subject protein plus the name of the organisms containing them, filling with the letter “M” all the other fields.

Figure 2.6 show an extract of the source code for the handling of multiple local alignments and for extracting the subject information from the BLAST output in case of single subject protein. The presence of multiple alignments is recognizes by the fact that the keyword “Score” is repeated more than one time in the same subject protein.

2.1.5 Post-processing

The “cvs” file produced for each organism in the data mining step described in the previous paragraph becomes the input file for the parametric analysis of protein amino acid sequence similarity. The parametric analysis of protein function identification and their annotation is performed in Microsoft Excel. Excel was chosen for its capability to show in efficient way tables of information and the capacity to use Visual Basic macro linked to the built-in functionality of Excel.

```
#############################################################################
#### After reading query information, start reading subject information
#############################################################################
while (<infile_all>) {

        $querylines .= $_;
        #### Check end of subject information
        if( $querylines =~/Lambda     K      H/){
                $counter_exit=$counter_exit+1;
                last;
        }
}

### Each new subject is recognized by ">ref"
my @sbjct_array = split (/\n>ref/,$querylines);
my $sbjct_array_size = @sbjct_array;

my $sbjct_index =0;
#### LOOP for all subjects
for ($sbjct_index = 1; $sbjct_index < $sbjct_array_size; $sbjct_index++) {

        $nameline = $sbjct_array[$sbjct_index];

        #### Read subject data
        $nameline =~ /^\|(\w+\.\d*)\|(.*)/;
        $sbjct_ref=$1;

        #### Avoid to report the case when query = subject
        if ($sbjct_ref=~$query_ref){ $discarded_hit_counter=$discarded_hit_counter+1;}
```

Figure 2.5. HYPE Data mining source code. Rules for retrieving subject protein information.

```
##########################################################################################
### Sometimes subject protein has multiple score (i.e. alignment) in this case this is reported
### in the output file
##########################################################################################
                my @sbjct_score_array = split (/Score =/,$nameline);
                my $sbjct_score_array_size = @sbjct_score_array;
                if ($sbjct_score_array_size >2) {
                        print outfile_res "$query_gi, $query_ref,$query_name,$query_type,
                        M,M,M,M,$sbjct_ref,$sbjct_name,$sbjct_type,M,M,M,M,M,
                        M,M,M,M,M,M,M,$sbjct_hyp_flag,$same_type_flag\n";
                }
                else{
    ### READ ALIGNMENT INFORMATION
                        $nameline =~ /Length=(\d+)/;
                        $sbjct_length =$1;
                        $nameline =~ /Score.*\s+(\d+(\.\d+)?) bits/;
                        $sbjct_score =$1;
                        $nameline =~ /Expect.*\s+(\d+(e-\d+)?(\.\d+)?), Meth/;
                        $sbjct_e=$1;
                        $nameline =~ /Identities =\s+(\d+)\/(\d+)/;
                        $sbjct_identities=$1;
                        $sbjct_identities_comp=$2;
                        $nameline =~ /Positives =\s+(\d+)\/(\d+)/;
                        $sbjct_pos=$1;
                        $sbjct_pos_comp=$2;
                        $nameline =~ /Gaps =\s+(\d+)\/(\d+)/;
                        $sbjct_gaps=$1;
                        $sbjct_gaps_comp=$2;
                        . . .
```

Figure 2.6. HYPE Data mining source code for the handling of multiple sequence alignment.

The generated Visual Basic function for the parametric analysis is base on the use of the Excel filter function. The parametric analysis can be performed by varying three criteria of similarity.

- **Minimal percentage of sequence similarity**. The minimal percentage of amino acid sequence similarity for showing the alignment results between query and subject protein. Percentage of sequence similarity refers to the number of identical amino acids plus positive amino acid substitutions with respect to the total alignment length.
- **Minimal percentage of query coverage.** The minimal percentage of query protein coverage for showing the alignment results between query and subject protein. Percentage of query coverage refers to the number of amino acids aligned with respect to the total protein length of the query protein.
- **Maximum query length**. The maximum number of amino acids composing the query protein for showing the alignment results between query and subject protein.

Other criteria of similarity applied to the values extracted from the data mining process and described in paragraph 2.1.4 could be easily introduced.

The post-process analysis can be performed on all the organisms reported in Table 1.1 or on a single organism by selecting it from an Excel “list box”. Figure 2.7 shows an extract of the Visual Basic macro that read the criteria of similarity and the data mining output file for the selected organism. The results of the parametric analysis can be reported in different Excel sheets to be easily compared. The results summarizing the HYPE annotation process are reported updating a table in a separate spreadsheet.

Two spreadsheets are generated for each organism. The first spreadsheet contains the list of hypothetical proteins that meet the criteria of similarity, the second spreadsheet groups together hypothetical proteins that align with proteins in multiple organisms.

HYPE post-process can be run with different type of configurations. The different configurations can be selected by changing the value of the Excel cell named "Hypothetical Protein Flag". There are four possible choices.

- **Hypothetical Protein Flag=0**: Subject proteins that are tagged "hypothetical" are not reported in the two output spreadsheets of HYPE. This configuration could be useful to consider in the annotation only proteins with a known function. Another use of this configuration is to highlight how the annotations were done for the selected organism and draw the attention to non standard annotation text such as "unknown protein", "undetermined" etc. This could provide the researcher with the hint to re-run the HYPE analysis with a different keyword with respect to "hypothetical" to analyze other uncharacterized protein.
- **Hypothetical Protein Flag=1**: Only subject proteins that are tagged "hypothetical" are reported in the two output spreadsheets of the HYPE. This configuration could be useful to analyze if a particular hypothetical protein is conserved and undefined in other organisms.
- **Hypothetical Protein Flag=2**: All subject proteins that meet the criteria of similarity defined by the user are reported in the two output spreadsheets of the HYPE independently from their name. This configuration is the one that gives the highest number similarity matches.

```
homology_perc = Range("B1").Value
querycov_perc = Range("B2").Value
querylength = Range("B3").Value
hypotheticalflag = Range("B4").Value
resultdifferentsheet = Range("B5").Value

test_ctr1 = ">" & Str$(homology_perc)
test_ctr2 = ">" & Str$(querycov_perc)
test_ctr3 = "<" & Str$(querylength)
test_ctr4 = Str$(hypotheticalflag)

title_txt = "Results for Percentage Similarity " & test_ctr1
title_txt = title_txt & " Percentage Query Coverage " & test_ctr2
title_txt = title_txt & " Query Length " & test_ctr3
title_txt = title_txt & " Hypothetical=" & test_ctr4

summary_number_of_homology = 0
summary_number_of_mult_homology = 0
summary_number_of_multitype_homology = 0
summary_number_annotations = 0
summary_number_multiple_annotation = 0

case_type = Range("E117").Value

data_worksheet_name = Range("B117").Value
data_workbook_name = data_worksheet_name & ".csv"


Workbooks.Open Filename:=data_workbook_name

' This instruction is needed for Mac Excel 2011
ActiveSheet.Name = data_worksheet_name
header_row = Sheets(data_worksheet_name).Rows("1:1").Value

If resultdifferentsheet <> 1 Then
   sheetname2 = "Results2"
   sheetname1 = "Results1"
Else
   sheetname2 = "Res2_" & case_type
   sheetname1 = "Res1_" & case_type
 End If

Windows(postprocess_workbookname).Activate
```

Figure 2.7. HYPE post-process interfacing with Excel spreadsheet.

- **Hypothetical Protein Flag>2**: This is the default configuration for proposing hypothetical protein annotation. The first output spreadsheet will have no subject protein with tag “hypothetical”. All the HYPE proposed annotations, in theory, do not refer to uncharacterized proteins. The second output spreadsheet shows all the similarity matches independently from the name of the subject protein. Because the second spreadsheet groups together hypothetical proteins that align with proteins in multiple organisms, this choice allows assessing with more confidence the conservation of a hypothetical protein in different organism.

Figure 2.8 shows part of the Visual Basic code that handles the above described HYPE execution configurations.

2.2 Assumptions and Limitations

HYPE differentiates an hypothetical protein from a protein of known function by the presence of the word “Hypothetical“. in the protein name. This approach does not guarantee to extract all the hypothetical proteins due to the fact that database authors do not consistently name all hypothetical proteins in the same way. The search of hypothetical protein annotated with different name is always possible, but requires a modification of the source code in the line highlighted in Figure 2.1. The same modification can be used to extend HYPE utilization to all uncharacterized proteins in a database.

BLAST searches were limited to an “e value” less than 0.0001, in order to avoid too many BLAST hits. This limitation is deemed not affecting the overall HYPE results,

```
   If hypotheticalflag < 2 Then
      Selection.AutoFilter
      ActiveSheet.Range("$A:$Y").AutoFilter Field:=5, Criteria1:=test_ctr2, Operator:=xlAnd
      ActiveSheet.Range("$A:$Y").AutoFilter Field:=7, Criteria1:=test_ctr1, Operator:=xlAnd
      ActiveSheet.Range("$A:$Y").AutoFilter Field:=22, Criteria1:=test_ctr3,
Operator:=xlAnd
      ActiveSheet.Range("$A:$Y").AutoFilter Field:=24, Criteria1:=test_ctr4
   Else
      Selection.AutoFilter
      ActiveSheet.Range("$A:$Y").AutoFilter Field:=5, Criteria1:=test_ctr2, _
         Operator:=xlAnd
      ActiveSheet.Range("$A:$Y").AutoFilter Field:=7, Criteria1:=test_ctr1, _
         Operator:=xlAnd
      ActiveSheet.Range("$A:$Y").AutoFilter Field:=22, Criteria1:=test_ctr3
   End If

   Dim rng As Range
   Set rng = ActiveSheet.AutoFilter.Range
   filterd_out_number =rng.Columns(1).SpecialCells(xlCellTypeVisible).Count-1
   all_out_number = rng.Rows.Count - 1
   copy_rng = "A1:Y" & Format(all_out_number + 1)
   rng.Range(copy_rng).SpecialCells(xlCellTypeVisible).Copy
   summary_number_of_homology = filterd_out_number

   If filterd_out_number > 0 Then
      'Range("A:Y").Select
      'Selection.Copy
      Windows(postprocess_workbookname).Activate
      Sheets(sheetname1).Select
      Range("A3").Select
      'rng.Range(copy_rng).SpecialCells(xlCellTypeVisible).PasteSpecial
Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
      'ActiveSheet.Paste
      Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False

      Application.DisplayAlerts = False
      Windows(data_workbook_name).Close
      Application.DisplayAlerts = True
      Workbooks(postprocess_workbookname).Sheets(sheetname1).Select
```

Figure 2.8. HYPE post-process. Different output result capability.

because the criteria of similarity required for the assignment of a function to a hypothetical protein are typically much more stringent.

The output of the blastp command is in the standard web-based BLAST output format; this allows recording the maximum amount of information regarding sequence alignments, but with the disadvantage of the necessity to develop data mining tools to manage large file. This approach avoids the risk to re-run BLAST in case additional information concerning the BLAST hits is needed in the follow-on analysis, but it requires the handling a huge amount of information. The execution of blastp for all the hypothetical proteins is the most time consuming part of HYPE running time.

HYPE does not collect data for proteins with multiple alignments; the main reason relies on the fact that the annotation could be different between two local alignments. Multiple local alignments may refer to multiple different domains within the same protein; the annotation in this case must be evaluated by a reviewer. When multiple alignments are found, HYPE reports only the protein name of the query and subject protein plus the name of the organisms containing them, filling with the letter “M” all the other fields.

Possible different approach to handle multiple local alignments could be to report the alignment with the maximum BLAST score, or to report all the alignments with a flag indicating that they are related to the same query protein. In the first case, the protein functional identification could be incorrect, in the second case a deeper analysis by the reviewer must occur anyway reducing the benefit of the automatic annotation proposed by HYPE.

CHAPTER 3

HYPE STRUCTURAL SIMILARITY

3.1 Materials and Methods

3.1.1 Search for Hypothetical Protein Structure

HYPE is capable of proposing annotations based on protein structural similarities using the NCBI web-based tool VAST. Hypothetical proteins for all the analyzed organisms in Table 1.1 were derived with the steps described in paragraphs 2.1.2 using RefSeq database created according to paragraph 2.1.1. The above procedures do not provide any information if a hypothetical protein had an experimentally determined structure. The only protein structure database searchable with BLAST is the Protein Data Bank (PDB) (Rose, Beran et al. 2011). PDB was simply downloaded from NCBI ftp server: “ftp://ftp.ncbi.nih.gov/blast/db/” and filename “pdbaa.tar.gz”. As of January 2012 PDB contains more than 78000 protein structure (RCSB 2012).

HYPE was used to search sequence similarity on PDB database with criteria of similarity equal to: percentage of sequence similarity equal to 90%, percentage of query and subject protein coverage equal to 90%. These criteria allow the association of protein structures present in PDB to more than one thousand hypothetical proteins

3.1.2 Protein Structural Similarity Search

Once protein structures in PDB were associated to the hypothetical proteins to be investigated, the structural similarity match was performed with NCBI VAST. VAST is a

web based tool that does not allow the possibility to evaluate the structural similarity automatically and recursively for all the structures associated to hypothetical proteins. HYPE is capable of running recursively VAST by modifying the webpage source code associated to VAST. Figure 3.1 illustrates the initial VAST webpage and the associated source code extracted using Microsoft Internet Explorer. The webpage source code is accessible using almost any web browser and can be modified by just editing it. Several procedures were developed during this work to automatically run VAST for all the hypothetical protein associated structures; the final solution was to use a Perl script.

The initial VAST webpage for a single protein can be called with the URL: "http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=PDBID", where PDBID is the four letter code identifying a structure in PDB database.

The VAST search can be executed by clicking on the button "VAST". The button loads the following second URL: "http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?sdid=InternalVASTCode", where InternalVASTCode is a unique code associated to the VAST search for the chosen PDB ID protein. For example, the protein structure with PDBID equal to "1XNE" has an associated InternalVASTCode of "126702". The InternalVASTCode can be retrieved by reading the source code of the initial VAST webpage. An extract of the HYPE source code for the recursive execution of VAST, as described above, is showed in Figure 3.2. The following steps are highlighted in Figure 3.2.

- Read from a file the PDBID of all hypothetical protein.
- Load the initial VAST webpage.
- Read the InternalVASTCode form the initial VAST webpage.

There can be multiple domains inside a protein or multiple structures; HYPE can handle all these types of situations. The correct execution of the structural similarity match in HYPE is highly dependent on the internal structure of the VAST webpage, any change in the webpage could have impact on the capability of HYPE to run recursively VAST on all the hypothetical protein structures to be analyzed.

After the VAST search is executed, a second part of HYPE source code collects the results from the NCBI webpage.

By default, VAST results are in graphical format, but they can be transformed in tables of values by selecting “Table” instead of “Graphics” in a pull down list in the webpage. This selection allows the loading of an additional URL containing a large series of parameters. The loading of this URL is highlighted in Figure 3.3. The internet address was generated by combining the InternalVASTCode with some constant parameters. The constant parameters are contained in the variables defined as “$text1” till “$text6” that are defined in a different part of HYPE source code and consequently not reported in Figure 3.3. The InternalVASTCodes are contained in the variable “list_html2get_nr_ids”, this is an array of VAST internal codes each representing a different domain in the query structure.

Once the VAST webpage with the results in table format is showed, HYPE reads the code of this new webpage to extract all the data and store them in a file.

The information is available after the html line:

“href="http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=”

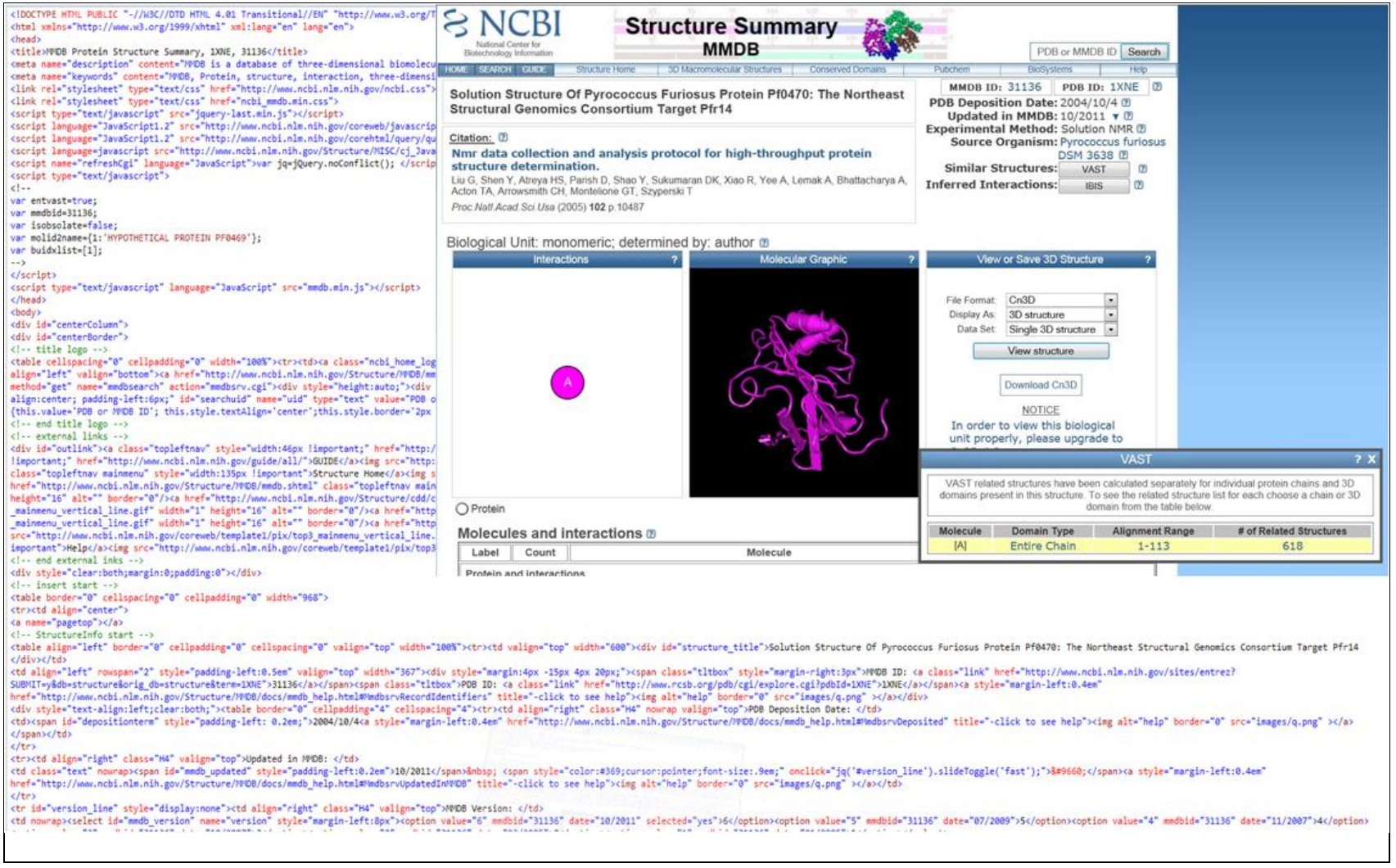


Figure 3.1. VAST webpage with associated source code.

```
##########################################################################################
################ FIND STRUCTURES WITH VAST RESULTS ######################################
##########################################################################################
while (<infile_all>) {      my $querylines =$_;
        my @tmp=split(/,/,$querylines);
        my $query_pdb=$tmp[8];

        if (length($query_pdb)==4){
                my $http_toget= $webpage_addr. $query_pdb;
                my $content = get($http_toget) or $flag_error=1;

                if ($flag_error==0){
                        my @hrefs = split(/href=\"/, $content);
                        my $hrefs_array_size = @hrefs;
                        my $i=0; my $j=0;
                        foreach (@hrefs) {

                                if (/www\.ncbi\.nlm\.nih\.gov\/Structure\/vast\/vastsrv\.cgi\?sdid=(\d+)/){
                                        $list_html2get_ids[$j]=$1;
                                        my @html2get = split(/\"/);
                                        $list_html2get[$j]=$html2get[0];
                                        #print "Sono qui $i $list_html2get[$j]   $list_html2get_ids[$j]\n";
                                        $j=$j+1;
                                }
                        }
                        # remove replications
                        my $remove_flag=0;     }
```

Figure 3.2. HYPE source code for the automatic execution of NCBI VAST.

The values collected refers to the first five best aligned structures, this choice is due primarily because is relatively rare to find a large number of structures that have high level of commonalities and secondary to avid a large amount of information to be retrieved. If a structure has several domains, the above described collection strategy is repeated for each domain.

A change of the VAST result webpage will likely require an update of the HYPE source code, however the HYPE source code should be unaffected if the webpage changes are limited only in the layout of the page and not in its content.

3.1.3 Post-Processing

The post-processing of VAST results is performed by a Microsoft Excel macro that reformats the information contained in the file produced by HYPE after the recursive execution of VAST. The goal of the post-processed file is to facilitate the identification of protein function. The file produced by HYPE is reported in Table 3.1; it is formed by thirty-four columns of data for a single hypothetical protein.

The data are divided in three sections; the first section highlighted in red has the most significant information about the query hypothetical protein and the similar protein found by a BLAST search on PDB database. It also contains the following major results of the amino sequence similarity match: percentage of similarity, query and subject protein coverage and percentage of alignment gaps. The second section showed with white background in Table 3.1 contains all the other parameters reported by HYPE after a sequence similarity match and described already in paragraph 2.1.4. This section is usually hidden to the reviewer to make easy the browsing of the VAST results.

```
        ##########################################################################################
        ######################################READ VAST RESULTS ##################################
        ##########################################################################################
                        for ($i = 0; $i < $non_redundant_html_count; $i++){
                                my $http_toget =$text1.$list_html2get_nr_ids[$i].$text2.$text3.$text4.$text5.$text6;
                                my $content = get($http_toget) or $flag_error=2;
                                if ($flag_error==0){
                                        my @hrefs = split(/class=\"Text\"\>\<a
href=\"http\:\/\/www\.ncbi\.nlm\.nih\.gov\/Structure\/mmdb\/mmdbsrv\.cgi\?uid=/, $content);

                                        my $hrefs_array_size = @hrefs;
                                        # Search only the first 5 more similar structures
                                        if ($hrefs_array_size> 6){$min_similar_struct=6;}
                                        else{$min_similar_struct=$hrefs_array_size;};
                                        for ($j = 1; $j < $min_similar_struct; $j++){
                                                if ($hrefs[$j] =~ /\"\>(.*?)\<\/a\> \<a/){
                                                        $sbjt_name=$1;
                                                        my @td = split(/\<td align=\"middle\" class=\"text\"\>\<font\>/, $hrefs[$j]);
                                                        my $td_array_size = @td;
                                                        if ($td[0] =~/0\"\>(.*)\<\/font/) {$ali_len =$1;} else{die "error in td0";}
                                                        if ($td[1] =~/(.*)\<\/font/) {$score=$1;} else{die "error in td1";}
                                                        if ($td[2] =~/(.*)\<\/font/) {$eval =$1;} else{die "error in td2";}
                                                        if ($td[3] =~/(.*)\<\/font/) {$rmsd =$1;} else{die "error in td3";}
                                                        if ($td[4] =~/(.*)\<\/font/) {$perc_id =$1;} else{die "error in td4";}
                                                        if ($td[5] =~/(.*)\<\/font/) {$date =$1;} else{die "error in td5";}
                                                        if ($td[6] =~/(.*)\<\/font/) {$lhm =$1;} else{die "error in td6";}
                                                        if ($td[7] =~/(.*)\<\/font/) {$gsp =$1;} else{die "error in td7";}
                                                        if ($td[8] =~/(.*)\<\/font/) {$sbjt_descr =$1;} else{die "error in td8";}
```

Figure 3.3. HYPE source code for collection of VAST results.

The third section highlighted in green reports the VAST results for the hypothetical protein as collected from the NCBI VAST webpage. The meaning of these VAST results is already described in the previous paragraph 1.6.2.

The post-processing functionality of HYPE operates in two different ways on the VAST output file. The first part removes the information with white background producing a post-processed file containing only the information highlighted in red and green in Table 3.1. It also removes all the results that contain in the “Subject Name” field the word “Hypothetical” and/or “Uncharacterized”.

Only hypothetical proteins that have a high amino acid sequence similarity to a protein already characterized in PDB will be shown. Independently from structural superimposition, the function of a hypothetical protein could already be derived just by the sequence similarity.

This is shown by the first result row of Table 3.2. HYPE search on PDB database reports that the hypothetical protein PF0537 has 99.8% amino acid sequence similarity with the argonaute protein 1Z25. Without running a VAST search, HYPE is proposing the annotation of hypothetical protein PF0537 as an Argonaute protein.

The second part of the post-processing functionality uses the VAST results to propose possible annotations. It operates in the opposite way with respect to the first part of the post processing. In fact, each hypothetical protein under investigation is considered only if it has a high amino acid sequence similarity to a protein not characterized in PDB. In this case, the VAST search will be used to find a structure from a characterized protein that is similar to the structure of the uncharacterized protein in PDB and consequently similar to the hypothetical protein under investigation. This is shown in Table 3.2. With

reference to the second result row, the hypothetical protein TK0174 does not have a determined structure, but it has an amino acid sequence similarity match of 93.7% with the protein 1VAJ in PDB. Assuming the amino acid difference does not have a big impact in the 3D protein structure; 1VAJ structure is compared to the other protein structures by a VAST search. The VAST results showed that 1T3N, a DNA polymerase protein, has a structural similar to 1VAJ and consequently to the hypothetical protein TK0174. For this reason, hypothetical protein TK0174 could have a function potentially similar to a “DNA polymerase”.

3.1 Assumptions and Limitations

HYPE was initially used to search sequence similarity on PDB database with criteria of similarity equal to: percentage of sequence similarity equal to 100%, percentage of query and subject protein coverage equal to 100%. These criteria allow the selection of protein structures in PDB that have amino acid sequence identical to the hypothetical protein under investigation and consequently that have an experimentally determined structure. In reality, HYPE considers 100% similar two proteins not only if they have identical amino acids but also if there are positive substitutions in the protein alignment. In this latter case, the found structure does not belong to the searched hypothetical protein but to a protein that is, nevertheless, very similar to the hypothetical protein according to the BLAST algorithms. Only twenty hypothetical proteins were found that satisfy the above stringent similarity criteria.

Table3.1. HYPE output file for a VAST search.

Query Gi	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percentage Gaps	Subject Ref	Subject Name	Subject Type	Subject Score	Subject Expected	Number of Subject Identities	Number of Subject Positives	Number of Subject Gaps	Number of Amino Acids Aligned	Subject Start	Subject End	Subject Length	Query Start	Query End	Query Length	Hypothetical Flag	Same Type Flag	PDB	Alignment Length	Score	E value	rmsd	Identity Percentage	Loop Hausdorff Metric	Gapped Score	Description
6E+07	YP_182	hypothe	Thermoc	100	99.115	90.1786	0	1XNE	A Chain A	NA	179	3.00E-46	85	101	0	112	1	112	113	1	112	112	0	0	2Z0T	93	13.2	10e-11.3	1.7	32.3	2.1	2	Crystal Structure Of Hypothetical Protein Ph0355
6E+07	YP_182	hypothe	Thermoc	100	99.115	90.1786	0	1XNE	A Chain A	NA	179	3.00E-46	85	101	0	112	1	112	113	1	112	112	0	0	2DP9	68	8	0.0062	2.1	16.2	5.7	3.4	Crystal Structure Of Conserved Hypothetical Protein Ttha0113 From Thermus Thermophilus
6E+07	YP_182	hypothe	Thermoc	100	99.115	90.1786	0	1XNE	A Chain A	NA	179	3.00E-46	85	101	0	112	1	112	113	1	112	112	0	0	1T62	67	10.7	10e-6.5	1.4	16.4	5.5	2.3	Crystal Structure Of Conserved Hypothetical Protein [gi:29377587] From Enterococcus
6E+07	YP_182	hypothe	Thermoc	100	99.115	90.1786	0	1XNE	A Chain A	NA	179	3.00E-46	85	101	0	112	1	112	113	1	112	112	0	0	2GBS	60	8.1	0.0051	1.8	8.3	7.9	3.4	Nmr Structure Of Rpa0253 From Rhodopseudomonas Palustris. Northeast
6E+07	YP_182	hypothe	Thermoc	100	99.115	90.1786	0	1XNE	A Chain A	NA	179	3.00E-46	85	101	0	112	1	112	113	1	112	112	0	0	1TE7	59	8.1	0.0019	2.2	20.3	4.4	4.2	Nmr Solution Structure Of The 14kda Hypothetical Protein Yqfb From Escherichia Coli
6E+07	YP_182	hypothe	Thermoc	100	95.794	93.6585	0	1VAJ	A Chain A	NA	355	6.00E-99	173	192	0	205	2	206	214	1	205	205	0	0	1ZQ7	137	11.9	10e-5.6	2.2	40.9	5.7	1.8	X-Ray Structure Of The Hypothetical Protein Q8pzk8 From Methanosarcina Mazei At The Resolution 2.1a. Northeast Structural Genomics
6E+07	YP_182	hypothe	Thermoc	100	95.794	93.6585	0	1VAJ	A Chain A	NA	355	6.00E-99	173	192	0	205	2	206	214	1	205	205	0	0	1WSC	135	10	0.0028	2.8	37.8	6.3	2.2	Crystal Structure Of St0229
6E+07	YP_182	hypothe	Thermoc	100	95.794	93.6585	0	1VAJ	A Chain A	NA	355	6.00E-99	173	192	0	205	2	206	214	1	205	205	0	0	1EZ1	40	6.4	0.043	2.3	2.5	8.2	6.5	Structure Of Escherichia Coli Purt-Encoded Glycinamide Ribonucleotide Transformylase
6E+07	YP_182	hypothe	Thermoc	100	95.794	93.6585	0	1VAJ	A Chain A	NA	355	6.00E-99	173	192	0	205	2	206	214	1	205	205	0	0	2J8U	38	6.7	0.027	2.2	7.9	18	6.4	Large Cdr3a Loop Alteration As A Function Of
6E+07	YP_182	hypothe	Thermoc	100	95.794	93.6585	0	1VAJ	A Chain A	NA	355	6.00E-99	173	192	0	205	2	206	214	1	205	205	0	0	1T3N	37	7.8	0.0402	2.4	5.4	7.2	7.4	Structure Of The Catalytic Core Of Dna Polymerase Iota In Complex With Dna And Dttp
6E+07	YP_182	hypothe	Thermoc	100	100	94.902	0	1V6T	A Chain A	NA	462	########	223	242	0	255	1	255	255	1	255	255	0	0	2DFA	241	35.1	10e-40.4	1.2	47.3	0.3	0.5	Crystal Structure Of Lactam Utilization Protein From Thermus Thermophilus Hb8
6E+07	YP_182	hypothe	Thermoc	100	100	94.902	0	1V6T	A Chain A	NA	462	########	223	242	0	255	1	255	255	1	255	255	0	0	2I5I	209	19.1	10e-12.7	3.3	15.8	9	1.7	Crystal Structure Of Hypothetical Protein (Ef3048) From Enterococcus Faecalis V583 At
6E+07	YP_182	hypothe	Thermoc	100	100	94.902	0	1V6T	A Chain A	NA	462	########	223	242	0	255	1	255	255	1	255	255	0	0	3NSN	203	16.2	10e-5.8	4.4	11.3	NA	NA	Crystal Structure Of Insect Beta-N-Acetyl-D-Hexosaminidase Ofhex1 Complexed With Tmg-
6E+07	YP_182	hypothe	Thermoc	100	100	94.902	0	1V6T	A Chain A	NA	462	########	223	242	0	255	1	255	255	1	255	255	0	0	1C7S	200	14.5	0.0147	4	8	11.9	2.2	Beta-N-Acetylhexosaminidase Mutant D539a Complexed With Di- N-Acetyl-Beta-D-
6E+07	YP_182	hypothe	Thermoc	100	100	94.902	0	1V6T	A Chain A	NA	462	########	223	242	0	255	1	255	255	1	255	255	0	0	2WHL	200	16	10e-6.1	4.6	8.5	8	2.4	Understanding How Diverse Mannanases Recognise Heterogeneous Substrates
6E+07	YP_182	hypothe	Thermoc	99.4	99.8	94.8	0.2	1UC2	B Chain B	NA	855	0	417	455	1	480	2	481	481	4	482	482	1	0	1UC2	480	60.6	10e-69.4	0.5	100	0	0.1	Hypothetical Extein Protein Of Ph1602 From
6E+07	YP_182	hypothe	Thermoc	99.4	99.8	94.8	0.2	1UC2	B Chain B	NA	855	0	417	455	1	480	2	481	481	4	482	482	1	0	1S6U	59	8.4	0.0153	2.4	5.1	17.3	4.4	Solution Structure And Backbone Dynamics Of The Cu(I) Form Of The Second Metal-Binding
6E+07	YP_182	hypothe	Thermoc	100	100	90.5882	0	2I0X	A Chain A	NA	139	3.00E-34	66	77	0	85	1	85	85	1	85	85	1	0	1FX2	69	6.9	0.0301	2.9	7.2	10.1	4.5	Structural Analysis Of Adenylate Cyclases From Trypanosoma Brucei In Their Monomeric State
6E+07	YP_182	hypothe	Thermoc	100	100	90.5882	0	2I0X	A Chain A	NA	139	3.00E-34	66	77	0	85	1	85	85	1	85	85	1	0	1U8S	66	7.5	10e-4.1	2.9	7.6	4.7	4.8	Crystal Structure Of Putative Glycine Cleavage System Transcriptional Repressor
6E+07	YP_182	hypothe	Thermoc	100	100	90.5882	0	2I0X	A Chain A	NA	139	3.00E-34	66	77	0	85	1	85	85	1	85	85	1	0	1ZPW	66	8.6	10e-7.0	1	25.8	0.4	1.7	Crystal Structure Of A Hypothetical Protein Tt1823 From Thermus Thermophilus
6E+07	YP_182	hypothe	Thermoc	100	100	90.5882	0	2I0X	A Chain A	NA	139	3.00E-34	66	77	0	85	1	85	85	1	85	85	1	0	3GMG	66	6.9	0.0097	2.6	4.5	8.2	4.3	Crystal Structure Of An Uncharacterized Conserved Protein From Mycobacterium
6E+07	YP_182	hypothe	Thermoc	100	100	90.5882	0	2I0X	A Chain A	NA	139	3.00E-34	66	77	0	85	1	85	85	1	85	85	1	0	3ONQ	66	6.3	0.0238	2.8	9.1	5.6	4.5	Crystal Structure Of Regulator Of Polyketide Synthase Expression Bad_0249 From

Table 3.2. Extract of HYPE report table of VAST search results. The result table can be divided in three sections. The first section represents the results of HYPE on PDB database for all the hypothetical proteins of one organism. The second section contains the corresponding protein structure ID and the corresponding annotation text. The third section contains the VAST structural similarity results extracted from the VAST webpage.

Hypothetical Protein				HYPE on PDB		HYPE VAST results			
Query Ref	**Query Name**	**Query Organism**	**Similarity [%]**	**PDB structure ID**	**Corresponding PDB structure Name**	**VAST Protein ID**	**VAST Score**	**VAST rmsd**	**Related Structure Name**
NP_578266.1	hypothetical protein PF0537	*Pyrococcus furiosus DSM 3638*	99.8	1Z25	A Chain A- Structure Of Pyrococcus Furiosus Argonaute With Bound Tungstate	NA	NA	NA	NA
YP_182468.1	hypothetical protein TK0174	*Thermococcus kodakarensis KOD1*	93.7	1VAJ	A Chain A- Crystal Structure Of Uncharacterized Protein Ph0010 From Pyrococcus Horikoshii	1T3N	7.8	2.4	Structure Of The Catalytic Core Of Dna Polymerase Iota In Complex With Dna And Dttp

The pool of available structures was increased by running HYPE on PDB as illustrated graphically in Figure 1.2 with the criteria of similarity of having amino acid sequence similarity match greater than 90% and query/subject protein coverage greater than 90%. With this configuration, the number of hypothetical proteins that align their amino acid sequence to proteins whose structure is defined in PDB increased from twenty to more than one thousand. The assumption is that the amino acids differences between the hypothetical protein under investigation and the corresponding aligned protein in PDB have a minimal impact on the protein 3D structure.

CHAPTER 4

RESULTS

4.1 Amino Acid Sequence Similarity Results

HYPE produces three tables of results when annotation of hypothetical proteins is based on amino acid sequence similarities. In the next subparagraphs, the results are reported, analyzed and explained.

4.1.1 HYPE Results: Summary Table

The first table is a summary report for all the organisms analyzed. Table 4.1 shows, as an example, the HYPE results when it is run with the following criteria of similarities: amino acid sequence similarities (see para. 2.1.4) between hypothetical proteins and any protein in RefSeq database greater than 98%, hypothetical protein coverage (see para. 2.1.4) greater than 90%. An explanation of the different values is as follows with reference to Table 4.1.

- **Number of hypothetical protein**: was computed by searching the keyword "hypothetical" in the protein annotation text.
- **Number of Hypothetical protein with similarity matches**: represents the number of proteins found by BLAST that satisfy the chosen criteria of similarities.
- **Number of Hypothetical Proteins with Multiple similarity matches**: indicates how many proteins, that satisfy the criteria of similarity, are present more than once in HYPE results.

- **Number of hypothetical proteins with multiple similarity matches across more than two organisms**: indicates if the multiple BLAST hits happen in two or more different organisms or different strains of the same organism.
- **Number of proposed annotation**: indicates how many annotations are proposed by HYPE. The annotation is proposed every time a BLAST hit does not contain the keyword “Hypothetical”.
- **Number of multiple annotations**: denotes the fact that HYPE produces more than one annotation for the same protein. This occurrence must be analyzed case by case by the user.

Considering as an example the row relative to *Thermococcus kodakarensis KOD1*, it can be observed that 61 of the 922 hypothetical proteins have a corresponding protein in RefSeq that satisfies the criteria of similarity selected. Thirteen of the 61 hypothetical proteins have multiple similarity matches and for all of them the similarities are in proteins present in different organisms (see last column of Table 4.6). HYPE proposed 8 annotations based on the fact that the BLAST hits do not contain the word “hypothetical” in the annotation text. One of the proposed 8 annotations has multiple annotation text that must be reviewed by the user.

Table 4.2 and Table 4.3 show additional two summary tables respectively for the criteria of sequence similarities greater than 90% and greater than 75%, maintaining constant the query and subject protein coverage to a value greater than 90%. The decrease of the level of required sequence similarity has the effect to increase the number of HYPE proposed annotations form 8 to 315 for the organism *Thermococcus kodakarensis KOD1*. Similar increases are observed in all the other organisms.

Table 4.1. HYPE output I. Results were produced by running HYPE with the criteria of having amino acid sequence similarities greater than 98% and hypothetical protein coverage greater than 90%.

Organism	Number of hypothetical proteins	Number of hypothetical proteins with similarity matches	Number of hypothetical proteins with multiple similarity matches	Number hypothetical proteins with multiple similarity matches across more than two organisms	Number of proposed annotations	Number of multiple annotations
Thermococcus kodakarensis KOD1	922	61	13	13	8	1
Pyrococcus furiosus DSM 3638	1017	79	21	19	13	6
Methanococcoides burtonii - DSM 6242	792	12	2	1	3	1
EcoliK12-sub MG1655	21	101	14	14	17	12
Saccharomyces cerevisiae S288c	995	1740	84	0	22	13
Magnaporthe oryzae	13769	7342	410	25	225	68
Arabidopsis thaliana	86	36	4	1	15	0
Populus trichocarpa	268	179	5	1	133	120
Danio rerio	5180	2111	257	20	1076	905
Caenorhabditis elegans	13824	3243	633	253	162	71
Mus musculus	3072	3592	438	109	1297	803
Homo Sapiens	2852	4632	972	635	2638	1632

Table 4.2. HYPE output II. Results were produced by running HYPE with the criteria of having amino acid sequence similarities greater than 90% and hypothetical protein coverage greater than 90%.

Organism	Number of hypothetical proteins	Number of hypothetical proteins with similarity matches	Number of hypothetical proteins with multiple similarity matches	Number hypothetical proteins with multiple similarity matches across more than two organisms	Number of proposed annotations	Number of multiple annotations
Thermococcus kodakarensis KOD1	69014	765	168	163	138	58
Pyrococcus furiosus DSM 3638	186497	1134	229	226	289	181
Methanococcoides burtonii - DSM 6242	259564	94	19	11	28	12
EcoliK12-sub MG1655	511145	161	16	16	23	17
Saccharomyces cerevisiae S288c	559292	2485	108	8	66	37
Magnaporthe oryzae	242507	15006	856	366	2839	2329
Arabidopsis thaliana	3702	97	5	3	28	10
Populus trichocarpa	3694	701	72	49	464	373
Danio rerio	8930	671	144	6098	5607	5606
Caenorhabditis elegans	6239	10858	3227	2626	861	531
Mus musculus	10090	12106	1011	446	5410	4386
Homo Sapiens	9606	11385	1607	1238	6843	5591

Table 4.3. HYPE output III. Results were produced by running HYPE with the criteria of having amino acid sequence similarities greater than 75% and hypothetical protein coverage greater than 90%.

Organism	Number of hypothetical proteins	Number of hypothetical proteins with similarity matches	Number of hypothetical proteins with multiple similarity matches	Number hypothetical proteins with multiple similarity matches across more than two organisms	Number of proposed annotations	Number of multiple annotations
Thermococcus kodakarensis KOD1	69014	2721	391	384	501	315
Pyrococcus furiosus DSM 3638	186497	4552	564	561	1770	1426
Methanococcoides burtonii - DSM 6242	259564	1838	217	200	418	302
EcoliK12-sub MG1655	511145	208	18	18	27	20
Saccharomyces cerevisiae S288c	559292	4456	238	124	645	480
Magnaporthe oryzae	242507	65053	3129	2551	32325	29924
Arabidopsis thaliana	3702	178	7	3	42	22
Populus trichocarpa	3694	3577	143	135	2376	2212
Danio rerio	7955	28994	1428	623	21972	20795
Caenorhabditis elegans	6239	36015	6147	5299	14703	13722
Mus musculus	10090	23855	1544	922	12970	11404
Homo Sapiens	*9606*	*18596*	*1751*	*1359*	*10532*	*9194*

4.1.2 HYPE Results: Proposed Annotation Table

For each organism in Table 1.1, HYPE produces a second result table containing detailed information about the proposed annotations. This second table, concerning *Thermococcus kodakarensis KOD1*, is partially reported in the following Table 4.4. The table was computed with the following criteria of similarities: amino acid sequence similarities between hypothetical proteins and any protein in RefSeq database greater than 98%, hypothetical protein coverage greater than 90%.

Each colunm has the following definition with reference to Table 4.4.

- **Query Ref**: represents the RefSeq accession number of the query protein.
- **Query name**: is the annotation text of the hypothetical protein on which the BLAST search is run.
- **Query cover [%]**: indicates the percentage of amino acids of the query protein that were aligned by the BLAST search with respect to the total number of amino acids forming the protein.
- **Subject Cover [%]**: is equivalent to “Query cover [%]” related to the BLAST hit (subject protein).
- **Similarity [%]**: indicates the percentage of amino acids aligned by BLAST. It includes identical amino acids plus positive substitutions (Korf, Yandell et al. 2003).
- **Gaps [%]**: shows the percentage of amino acids not aligned by the BLAST search.
- **Subject Ref**: is the RefSeq accession number of the BLAST hit.
- **Subject Name**: is the annotation text of the subject protein.

- **Subject Organism**: is the organism name containing the subject protein.

The HYPE results for all the twelve analyzed organisms are available in appendix for amino acid sequence similarity greater than 98% and query/subject protein coverage greater that 98%. With these criteria of similarities, HYPE proposes 474 annotations.

The" hypothetical protein TK0055" has 98.2% amino acid sequence similarity with "ProFAR isomerase associated superfamily protein" from *Thermococcus sp. AM4* (Table 4.4, first row).

100% Of amino acids have been aligned by BLAST, and there are no gaps in amino acids sequence similarity match. The last two rows show an example of multiple annotations. The hypothetical protein TK1693 was found to be similar to two proteins from two different organisms. One of the proposed annotations is, in this case, not particular helpful because it is annotated with the generic text:"conserved protein domain" and does not provide any hint of the possible function. It is reported by HYPE because it does not contain the word "hypothetical" in the annotation text. It is up to the reviewer to analyze case by case the significance of the HYPE proposed annotation.

A particular interesting case is when the criteria of similarities are chosen to be: amino acid sequence similarity equal to 100% and hypothetical protein coverage equal to 100%. HYPE found 54 hypothetical proteins in the twelve organisms analyzed that have a perfectly and completely aligned protein in another organism whose function was already characterized. These HYPE proposed annotations could be immediately used to update the RefSeq database by suggesting them to NCBI. Table 4.5 shows an extract of the HYPE results for this particular case.

Table 4.4. Annotations proposed by HYPE for *Thermococcus kodakarensis KOD1* that satisfy the criteria of similarity greater than 98% amino acid sequence similarities with greater than 90% of hypothetical protein coverage.

Query Ref	Query Name	Query Cover [%]	Subject Cover [%]	Similarity [%]	Gaps [%]	Subject Ref	Subject Name	Subject Organism
YP_182468.1	hypothetical protein TK0055	100.0	100.0	98.2	0.0	ZP_04878637.1	ProFAR isomerase associated superfamily protein	*Thermococcus sp. AM4*
YP_182771.1	hypothetical protein TK0358	99.4	100.0	98.5	0.0	YP_002307099.1	RNA terminal phosphate cyclase	*Thermococcus onnurineus NA1*
YP_183636.1	hypothetical protein TK1223	100.0	100.0	98.9	0.0	ZP_04878882.1	membrane bound hydrogenase- MbxD subunit	*Thermococcus sp. AM4*
YP_183691.1	hypothetical protein TK1278	100.0	100.0	99.1	0.0	ZP_04880486.1	DNA-binding protein	*Thermococcus sp. AM4*
YP_183974.1	hypothetical protein TK1561	99.0	98.5	98.0	0.0	YP_002958671.1	Multiple antibiotic resistance (Mar)-related protein	*Thermococcus gammatolerans EJ3*
YP_184003.1	hypothetical protein TK1590	100.0	100.0	98.8	0.0	ZP_04879917.1	RecA-superfamily ATPase implicated in signal transduction	*Thermococcus sp. AM4*
YP_184106.1	hypothetical protein TK1693	91.5	91.5	98.8	0.6	YP_002958739.1	Component of ring hydroxylating complex- putative	*Thermococcus gammatolerans EJ3*
YP_184106.1	hypothetical protein TK1693	91.5	91.5	98.8	0.6	ZP_04880559.1	conserved domain protein	*Thermococcus sp. AM4*

Table 4.5. HYPE proposed annotations in case of perfect and complete BLAST alignment. Perfect and complete alignment of protein is obtained by the criteria of similarity that amino acid sequence similarities are equal to 100%, hypothetical protein coverage is equal to 100% and subject protein coverage is equal to 100%.

Query Ref	Query Name	Query Organism	Query Cover [%]	Subject Cover [%]	Similarity [%]	Gaps [%]	Subject Ref	Subject Name	Subject Organism
NP_877907.1	hypothetical protein PF0785.3n	*Pyrococcus furiosus DSM 3638*	100	100	100	0	NP_877895.1	transposase	*Pyrococcus furiosus DSM 3638*
YP_001165334.1	hypothetical protein b4622	*Escherichia coli str. K-12 substr. MG1655*	100	100	100	0	YP_001746475.1	formate dehydrogenase H	*Escherichia coli TA280*
XP_001522527.1	hypothetical protein MGCH7_ch7g630	*Magnaporthe oryzae 70-15*	100	100	100	0	XP_965758.1	40S ribosomal protein S27	*Neurospora crassa OR74A*
XP_001522527.1	hypothetical protein MGCH7_ch7g630	*Magnaporthe oryzae 70-15*	100	100	100	0	XP_366796.1	40S ribosomal protein S27	*Verticillium albo-atrum VaMs.102*
NP_001070117.1	hypothetical protein LOC767711	*Danio rerio*	100	100	100	0	NP_001007786.1	crystallin- gamma M1	*Danio rerio*
XP_001335379.1	PREDICTED: hypothetical protein LOC796596	*Danio rerio*	100	100	100	0	NP_001107105.1	zinc finger-like gene 1	*Danio rerio*
NP_001103752.1	hypothetical protein LOC792137	*Danio rerio*	100	100	100	0	NP_919360.1	carbonyl reductase [NADPH] 1	*Danio rerio*
NP_001116790.1	hypothetical protein LOC799807	*Danio rerio*	100	100	100	0	NP_001038328.1	crystallin- gamma M2d5	*Danio rerio*

4.1.3 HYPE Results: Inter-Organism Similarity Table

The third type of results produced by HYPE is shown in Table 4.6. It groups together all the sequence similarity matches for each single hypothetical protein that satisfy the user defined similarity criteria. This table could be useful to highlight if a particular hypothetical protein is conserved among different organisms or even among different domains of life. This information could also be useful for running phylogenic profiles using multiple alignment sequence tool such as ClustalW2 (Chenna, Sugawara et al. 2003).

In the example shown in Table 4.6, all the hypothetical proteins are conserved among different organisms but the organisms are part of the same kingdom of life. HYPE post-process file can easily be used to evaluate which proteins in the analyzed organism are similar to proteins belonging to a specific organism by using the built-in filter function of Microsoft Excel on the field “Subject Organism”.

Table 4.6 gives also an indication if there are other proteins in different organisms that could be annotated with the same text. For example, considering the first two rows of Table 4.6, the “hypothetical protein TK0358” has a HYPE proposed annotation as “RNA terminal phosphate cyclase”. Because”hypothetical protein TGAM_1505” from *Thermococcus gammatolerans EJ3* shares great similarities with TK0358, it is likely that it can be annotated in the same way. However, annotation of TGAM_1505 could be definitely assessed only if HYPE would be run for the organism containing the protein TGAM_1505 i.e., “*thermococcus gammatolerans EJ3*”.

Table 4.6. BLAST hits for the same hypothetical protein that satisfies the criteria of similarity selected by the user that are grouped together. Protein conserved among multiple organisms can be revealed.

Query Ref	Query Name	Query Organism	Query Cover [%]	Subject Cover [%]	Similarity [%]	Gaps [%]	Subject Ref	Subject Name	Subject Organism
YP_182771.1	hypothetical protein TK0358	*Thermococcus kodakarensis KOD1*	99.4	100.0	98.5	0.0	YP_002959871.1	hypothetical protein TGAM_1505	*Thermococcus gammatolerans EJ3*
YP_182771.1	hypothetical protein TK0358	*Thermococcus kodakarensis KOD1*	99.4	100.0	98.5	0.0	YP_002307099.1	RNA terminal phosphate cyclase	*Thermococcus onnurineus NA1*
YP_183262.1	hypothetical protein TK0849	*Thermococcus kodakarensis KOD1*	100.0	100.0	98.0	0.0	NP_877771.1	hypothetical protein PH0630.1n	*Pyrococcus horikoshii OT3*
YP_183262.1	hypothetical protein TK0849	*Thermococcus kodakarensis KOD1*	100.0	100.0	98.0	0.0	YP_002995334.1	hypothetical protein TSIB_1936	*Thermococcus sibiricus MM 739*
YP_183262.1	hypothetical protein TK0849	*Thermococcus kodakarensis KOD1*	100.0	86.4	100.0	0.0	YP_002307630.1	hypothetical protein TON_1245	*Thermococcus onnurineus NA1*
YP_183691.1	hypothetical protein TK1278	*Thermococcus kodakarensis KOD1*	100.0	100.0	99.1	0.0	YP_002959562.1	hypothetical protein TGAM_1196	*Thermococcus gammatolerans EJ3*
YP_183691.1	hypothetical protein TK1278	*Thermococcus kodakarensis KOD1*	100.0	100.0	99.1	0.0	ZP_04880486.1	DNA-binding protein	*Thermococcus sp. AM4*
YP_183974.1	hypothetical protein TK1561	*Thermococcus kodakarensis KOD1*	99.0	98.5	98.0	0.0	YP_002958671.1	Multiple antibiotic resistance (Mar)-related protein	*Thermococcus gammatolerans EJ3*
YP_183974.1	hypothetical protein TK1561	*Thermococcus kodakarensis KOD1*	100.0	99.5	98.0	0.0	ZP_04879923.1	conserved hypothetical protein TIGR00427	*Thermococcus sp. AM4*

The same situation is present in other cases demonstrating the internal inconsistency of some RefSeq records; the utilization of a tool like HYPE by a database curator could highly increase the ability to resolve some of those issues.

4.2 Quality of annotation, a proposed approach

The number of proposed annotations of HYPE depends on the criteria of similarity defined by the user. Figure 4.1 shows the percentage of HYPE proposed annotations with respect to the total number of hypothetical proteins as function of the minimum percentage of amino acid sequence similarity.

Depending on the organism, it is theoretically possible to annotate more than 30% of hypothetical proteins if the level of similarities is chosen to be greater than 60%. This allows the annotation of a large number of hypothetical proteins, but unfortunately with a lower level of confidence. It is our opinion that this information is anyhow more useful for a researcher than a generic reference to an uncharacterized hypothetical protein.

An indication of the quality of annotation could be introduced in the annotation text to inform researchers of the criteria chosen for the annotation. Several possibilities exist, for example: dividing the quality of annotation in level of similarities (e.g. high, medium, low), or just reporting the key similarity parameters in the annotation text.

Table 4.7 shows this latter approach applied to HYPE proposed annotation text. The quality of annotation is introduced in the square brackets and includes four parameters: “A” indicating the percentage of similarity in the amino acid sequence alignment, “Q” and “S” for the percentage of protein coverage respectively for query and subject protein, and “G” for percentage of gaps in protein alignment. .

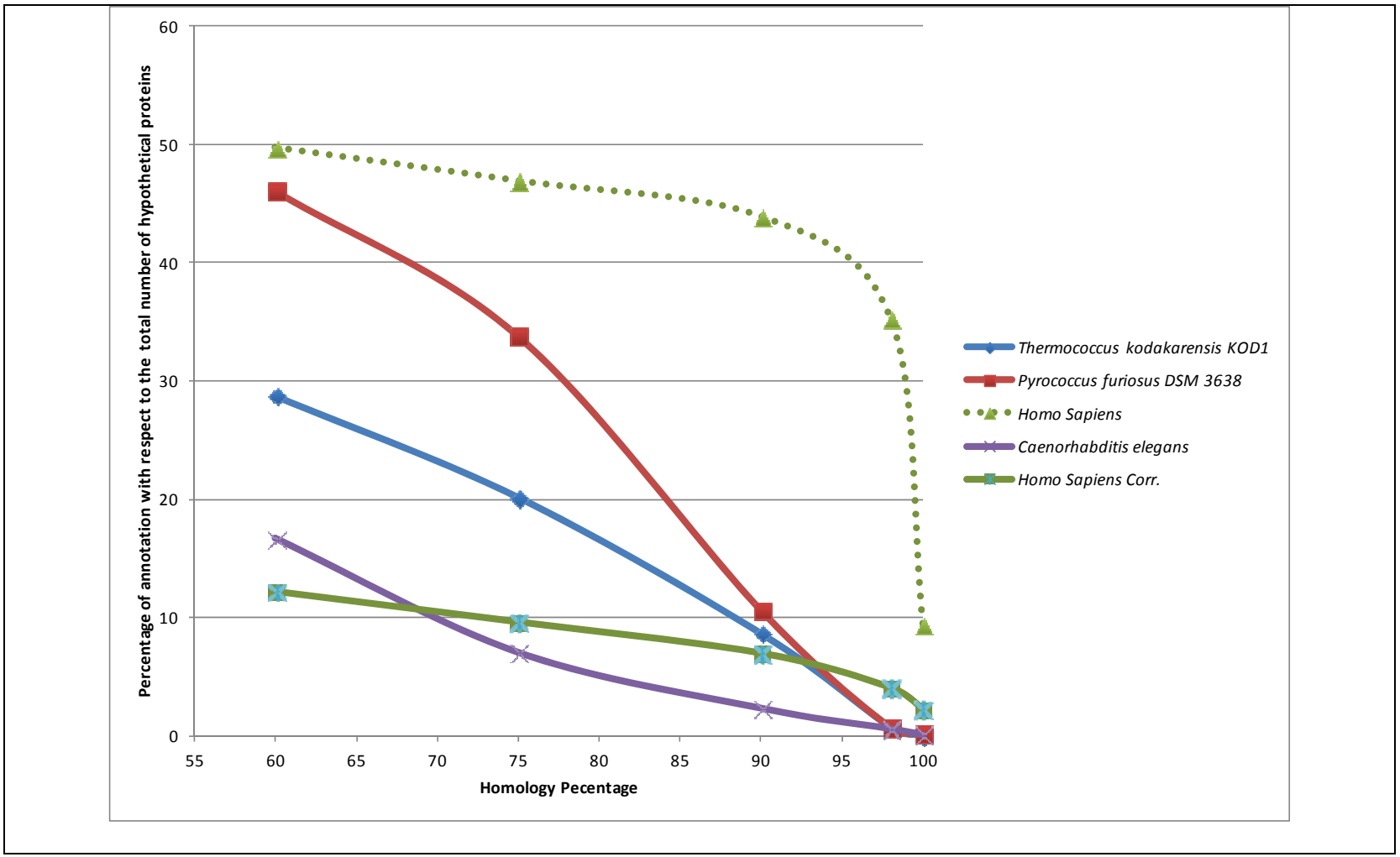


Figure 4.1. Parametric analysis of the HYPE proposed annotations as function of minimum amino acid sequence similarity. The values on the y-axes represent, for each organism, the percentage of proposed annotations with respect to the total number of hypothetical proteins. The level of hypothetical protein coverage is fixed to a value greater than 90% for all studied cases.

Table 4.7. Example of quality of annotation in HYPE proposed annotation text. The quality of annotation in red includes: the percentage of similarity in the amino acid sequence alignment preceded by the letter "A", the percentages of protein coverage for query and subject protein respectively preceded by the letters "Q" and "S", and the percentage of alignment gaps preceded by the letter "G".

RefSeq Name	Organism	Current HYPE Proposed Annotation	Example of Quality of Annotation
hypothetical protein TK0358	*Thermococcus kodakarensis KOD1*	RNA terminal phosphate cyclase	RNA terminal phosphate cyclase, [A98.5, Q99.4, S100, G0]
hypothetical protein b4622	*Escherichia coli str. K-12 substr. MG1655*	formate dehydrogenase H	formate dehydrogenase H, [A100, Q100, S100, G0]
hypothetical protein PF0009	*Pyrococcus furiosus DSM 3638*	thiazole biosynthesis adenylyltransferase	thiazole biosynthesis adenylyltransferase, [A66.8, Q92.2, S95.2, G4.1]

"

It will be immediately clear to a researcher fto the annotation text that the annotation for hypothetical protein b4622 as “formate dehydrogenase H” has a high level of accuracy even if it was derived from an automatic annotation tool; vice versa the annotation of hypothetical protein PF0009 as “thiazole biosynthesis adenylyltransferase” should be considered just as an indication of the possible protein function The indication of the quality of annotation increases the utility of tool such as HYPE by augmenting the number of undefined proteins that could be characterized while maintaining an accurate annotation text.

4.3 BLAST Scoring and Protein Identity

An interesting result noted during the identification of function for hypothetical protein through HYPE search is reported in Figure 4.2. The BLAST search for hypothetical protein with GI number equal to 198282015 gave two BLAST hits. The first hit was related to protein with RefSeq accession number equal to XP_002666347.2, the second hit was related to three identical proteins: NP_001103581.1, NP_001153834.1, NP_001153837.1. In this last case, the number of identical amino acids is 128/128, where 128 is the total number of amino acids of both the query and subject protein. This means that the query protein and subject protein are exactly the same.

In case of protein XP_002666347.2, the number of identical amino acids is still 128/128, but the subject protein length is of 186 amino acids. This means that the subject protein is identical to the query protein for only 128 amino acids, but has 58 more amino acids. This hit received a BLAST score of 264, while the hit to the proteins that were exactly the same received a score of only 261.

BLAST algorithms, for unknown reason, gave a higher score to the match between different proteins than to the match between two proteins that are exactly the same. These BLAST results, anyhow, do not have any consequences on the capacity of the tool to find similar protein.

A second unexpected behavior, this time related to the RefSeq database is highlighted in light blue in Figure 4.2. The proteins NP_001103581.1, NP_001153834.1, NP_001153837.1 are identical, they just belong to different loci in a chromosome, but they are annotated in different ways. In particular NP_001153837.1 has a function that has been characterized and annotated as “Novel rhamnose binding lectin-like”, but the other identical proteins are still annotated as “Hypothetical Protein”.

HYPE, in this case, does not collect all the different names, but only the first one. This choice could affect the final annotation results, because, even if all the three hypothetical proteins will be analyzed independently by HYPE, each time, HYPE could skip to collect the “valid” annotation just because is not the first one to appear. The source of the problem is related to the RefSeq database; anyhow a new version of HYPE will be developed to handle this kind of events.

4.4 Protein Structural Similarity Results

HYPE is capable of proposing annotations based on protein structural similarities using the NCBI web-based tool VAST. Hypothetical proteins with an experimental determined structure were initially found by a HYPE search on PDB database with criteria of similarity equal to 100%. This criterion allows the selection of protein structure in PDB that perfectly align their amino acid sequence to the hypothetical

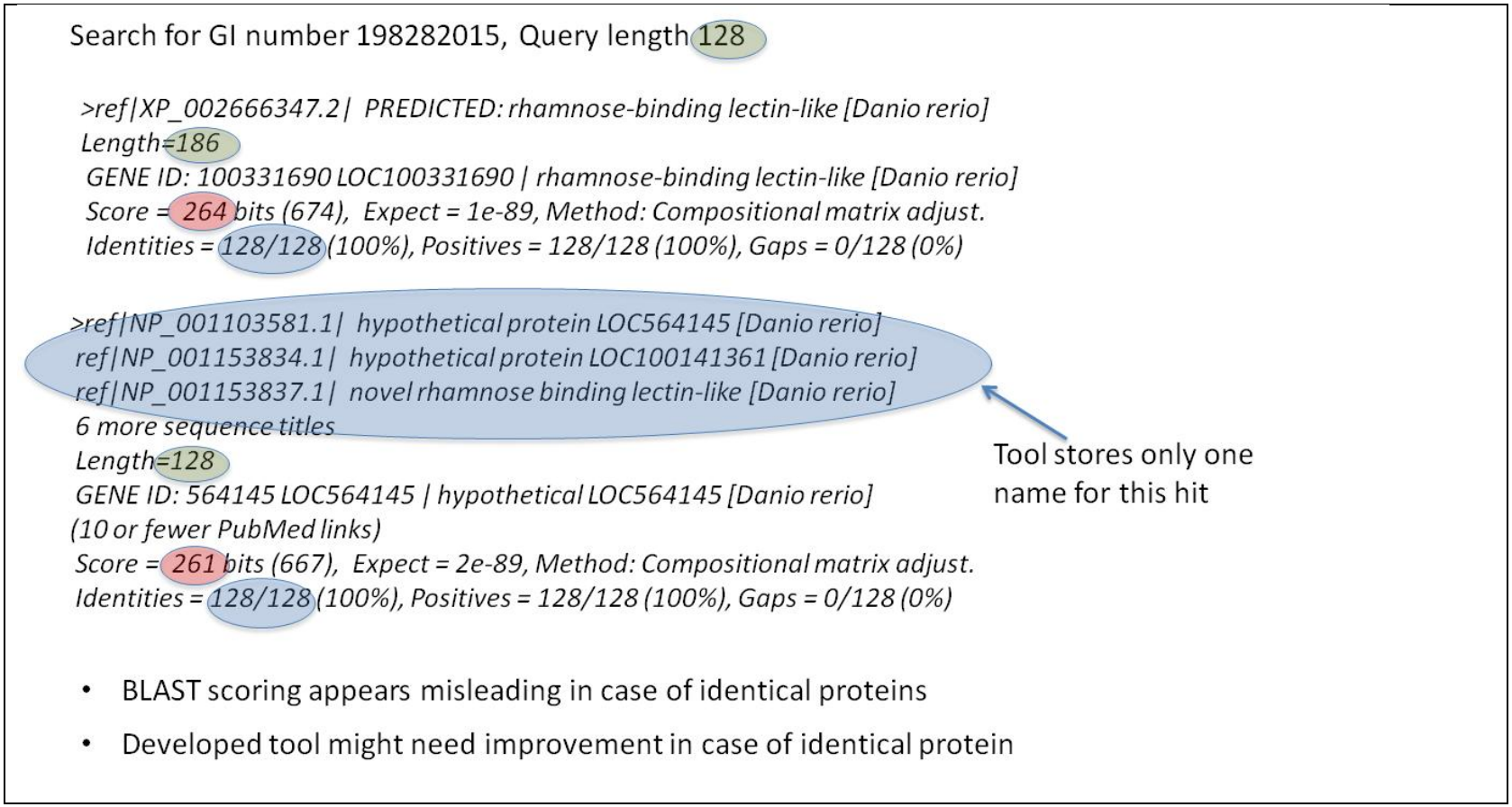


Figure 4.2. BLAST search on identical protein.

protein to be annotated. Only twenty of more than 40 thousand hypothetical proteins present in the twelve analyzed organisms were found to have an associated structure according the above mentioned criteria. In order to increase the number of hypothetical proteins with an associated structure, HYPE was run on PDB as illustrated graphically in Figure 1.2 with the following criteria of similarity: amino acid sequence similarity match greater than 90%; query/subject protein coverage greater than 90%. The number of hypothetical proteins with an associated structure increased from twenty to more than one thousand.

This augmented pool of protein structures can be used under the assumption that the amino acids differences between the hypothetical protein under investigation and the corresponding aligned protein in PDB have a minimal impact on the protein 3D structure.

The results of the VAST analysis are showed in the next two Table 4.8 and Table 4.9. As already described in paragraph 3.1.3, VAST results can be divided in two types. The first type of results is obtained by amino acid sequence similarity toward the PDB database. If the protein in PDB is already characterized, there is no need to consider also the structural similarity because it is already possible to annotate the hypothetical protein under investigation. This is illustrated in Table 4.8 where, for example, hypothetical protein PF0537 (first row) has 99.9 percent sequence similarity with an “argonaute bound to tungstate” protein from *Pyrococcus Furiousus*.

Independently from the structural analysis (highlighted in green), the hypothetical protein protein PF0537 can be annotated with high level of confidence as an “argonaute” protein.

Table 4.8. Protein function identification by sequence similarity match with protein in PDB database.

Query Gi	Query Ref	Query Name	Query Type	Percent Query Coverage	Percent Subject Coverage	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	PDB	Align Length	Score	E value	rmsd	Identity Percent	Loop Hausdorff Metric	Gapped Score	Description
18976909	NP_578266.1	hypothetical protein PF0537	Pyrococcus furiosus DSM 3638	100.0	99.9	99.9	0	1Z25	A Chain A- Structure Of Pyrococcus Furiosus Argonaute With Bound Tungstate	1Z26	192	16	10e-9.9	0.3	100	0	0.2	Structure Of Pyrococcus Furiosus Argonaute With Bound Tungstate
18976909	NP_578266.1	hypothetical protein PF0537	Pyrococcus furiosus DSM 3638	99.9	99.7	97.7	0	1U04	A Chain A- Crystal Structure Of Full Length Argonaute From Pyrococcus Furiosus	1Z26	181	16	10e-10.1	0.3	97.8	0	0.2	Structure Of Pyrococcus Furiosus Argonaute With Bound Tungstate
18976963	NP_578320.1	hypothetical protein PF0591	Pyrococcus furiosus DSM 3638	99.7	99.7	92.1	0	2ZU7	B Chain B- Crystal Structure Of Mannosyl-3-Phosphoglycerate Synthase From Pyrococcus Horikoshii	3O3P	249	21.8	10e-13.6	3.6	14.9	NA	NA	Crystal Structure Of R. Xylanophilus Mpgs In Complex With Gdp-Mannose
18976974	NP_578331.1	hypothetical protein PF0602	Pyrococcus furiosus DSM 3638	99.4	89.5	90.1	0	2R6V	A Chain A- Crystal Structure Of Fmn-Binding Protein (Np_142786.1) From Pyrococcus Horikoshii At 1.35 A Resolution	3BPK	168	22.4	10e-19.2	1.9	19	2.4	1.2	Crystal Structure Of Nitrilotriacetate Monooxygenase Component B From Bacillus Cereus
18977040	NP_578397.1	hypothetical protein PF0668	Pyrococcus furiosus DSM 3638	100.0	100.0	96.8	0	2DYY	L Chain L- Crystal Structure Of Putative Translation Initiation Inhibitor Ph0854 From Pyrococcus Horikoshii	1X25	125	15.7	10e-16.3	0.9	67.2	0.4	0.7	Crystal Structure Of A Member Of Yjgf Family From Sulfolobus Tokodaii (St0811)
18977389	NP_578746.1	hypothetical protein PF1017	Pyrococcus furiosus DSM 3638	100.0	100.0	93.5	0	3IVZ	B Chain B- Crystal Structure Of Hyperthermophilic Nitrilase	1J31	261	29.7	10e-34.3	0.5	84.3	0	0.2	Crystal Structure Of Hypothetical Protein Ph0642 From Pyrococcus Horikoshii
18977483	NP_578840.1	hypothetical protein PF1111	Pyrococcus furiosus DSM 3638	99.7	93.3	99.1	0	2QM3	A Chain A- Crystal Structure Of A Predicted Methyltransferase From Pyrococcus Furiosus	3S1S	191	16.8	10e-4.7	4	9.9	NA	NA	Characterization And Crystal Structure Of The Type Iig Restriction Endonuclease Bpusi
18977490	NP_578847.1	hypothetical protein PF1118	Pyrococcus furiosus DSM 3638	100.0	100.0	93.8	0	3PV9	E Chain E- Structure Of Ph1245- A Cas1 From Pyrococcus Horikoshii	2YZS	205	13.9	10e-9.2	1.3	50.7	NA	NA	Crystal Structure Of Uncharacterized Conserved Protein From Aquifex Aeolicus
18977503	NP_578860.1	hypothetical protein PF1131	Pyrococcus furiosus DSM 3638	99.6	97.8	100.0	0	3PKM	X Chain X- Crystal Structure Of Cas6 With Its Substrate Rna	3PKM	126	13.4	10e-8.9	0.7	100	0	0.6	Crystal Structure Of Cas6 With Its Substrate Rna
18977503	NP_578860.1	hypothetical protein PF1131	Pyrococcus furiosus DSM 3638	99.6	96.3	100.0	0	3I4H	X Chain X- Crystal Structure Of Cas6 In Pyrococcus Furiosus	3PKM	119	11.4	10e-6.4	0.9	100	NA	0.8	Crystal Structure Of Cas6 With Its Substrate Rna
18977563	NP_578920.1	hypothetical protein PF1191	Pyrococcus furiosus DSM 3638	97.7	73.9	100.0	0	2E0Z	C Chain C- Crystal Structure Of Virus-Like Particle From Pyrococcus Furiosus	2E0Z	235	29.2	10e-30.5	0.9	100	0.3	0.4	Crystal Structure Of Virus-Like Particle From Pyrococcus Furiosus
18977644	NP_579001.1	hypothetical protein PF1272	Pyrococcus furiosus DSM 3638	100.0	100.0	95.7	0	1V6T	A Chain A- Crystal Structure Of Lactam Utilization Protein From Pyrococcus Horikoshii Ot3	2DFA	241	35.1	10e-40.4	1.2	47.3	0.3	0.5	Crystal Structure Of Lactam Utilization Protein From Thermus Thermophilus Hb8
18977875	NP_579232.1	hypothetical protein PF1503	Pyrococcus furiosus DSM 3638	100.0	100.0	100.0	0	1GEF	B Chain B- Crystal Structure Of The Archaeal Holliday Junction Resolvase Hjc From Pyrococcus Furiosus Form Ii	1GEF	116	13.1	10e-11.0	1	100	0	0.9	Crystal Structure Of The Archaeal Holliday Junction Resolvase Hjc
18978303	NP_579660.1	hypothetical protein PF1931	Pyrococcus furiosus DSM 3638	100.0	100.0	98.8	0	2DR3	F Chain F- Crystal Structure Of Reca Superfamily Atpase Ph0284 From Pyrococcus Horikoshii Ot3	2DR3	229	23	10e-26.4	0.4	100	0.4	0.2	Crystal Structure Of Reca Superfamily Atpase Ph0284 From Pyrococcus Horikoshii Ot3

The second type of annotation results is based on structural similarities. This is run when also the protein with structure in PDB is uncharacterized. In other words, the hypothetical protein under investigation is similar in sequence to an uncharacterized protein structure in PDB. This structure is matched with all the other structure in PDB with a VAST search. The results of the VAST search are further filtered to show only matches to functionally defined proteins.

Hypothetical protein YJL068C from *Saccharomyces cerevisiae S288c* is highly similar in sequence to the yeast hypothetical protein Yjg8_yeast (Table 4.9, first row) that has a structure with PDB identifier equal to 1PV1. This structure is similar, according to VAST algorithm, to another structure 3I6Y that is annotated as “Esterase From The Oil-Degrading Bacterium Oleispira Antarctica”. It is likely that the initial hypothetical protein YJL068C from *Saccharomyces cerevisiae S288c* has the function of an esterase.

HYPE searches on protein structural similarity require a more detailed analysis of the results and, in general, more caution in the annotation. The reasons are multiple: the assumption concerning the similarity of the protein 3D structure in presence of amino acid sequence differences needs to be verified; proteins with similar structures could have a completely different active site and consequently have different functions; VAST score and root mean square of the distance between amino acid backbones are indication of structural similarities but they are not measure of the electrical and chemical proprieties of the proteins.

Table 4.9. Protein function identification by structural match to already characterized protein.

Query Gi	Query Ref	Query Name	Query Type	Percent Query Coverage	Percent Subject Coverage	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	PDB	Align Length	Score	E value	rmsd	Identity Percent	Loop Hausdorff Metric	Gapped Score	Description
18976623	NP_577980.1	hypothetical protein PF0251	Pyrococcus furiosus DSM 3638	99.2	99.7	95.7	0	2AS0	B Chain B- Crystal Structure Of Ph1915 (Apc 5817): A Hypothetical Rna Methyltransferase	1WXX	379	46.1	10e-43.8	2.3	39.3	1.5	0.7	Crystal Structure Of Tt1595
17550248	NP_509242.1	hypothetical protein C07D8.6	Caenorhabditis elegans	100.0	100.0	100.0	0	1QWK	A Chain A- Structural Genomics Of Caenorhabditis Elegans: Hypothetical 35.2 Kda Protein (Aldose Reductase Family Member)	3BCJ	293	32.9	10e-35.7	1.6	41	2	0.6	Crystal Structure Of Aldose Reductase Complexed With 2s4r (Stereoisomer Of Fidarestat
17550248	NP_509242.1	hypothetical protein C07D8.6	Caenorhabditis elegans	100.0	100.0	100.0	0	1QWK	A Chain A- Structural Genomics Of Caenorhabditis Elegans: Hypothetical 35.2 Kda Protein (Aldose Reductase Family Member)	1LQA	258	31	10e-30.4	2.5	24	4.9	1	Tas Protein From Escherichia Coli In Complex With Nadph
17550248	NP_509242.1	hypothetical protein C07D8.6	Caenorhabditis elegans	100.0	100.0	100.0	0	1QWK	A Chain A- Structural Genomics Of Caenorhabditis Elegans: Hypothetical 35.2 Kda Protein (Aldose Reductase Family Member)	1PZ1	262	28.4	10e-25.3	2.8	27.5	5.7	1.1	Structure Of Nadph-Dependent Family 11 Aldo-Keto Reductase Akr11b(Holo)
18976623	NP_577980.1	hypothetical protein PF0251	Pyrococcus furiosus DSM 3638	99.2	99.7	95.7	0	2AS0	B Chain B- Crystal Structure Of Ph1915 (Apc 5817): A Hypothetical Rna Methyltransferase	2IGT	234	28.1	10e-15.8	2.5	19.2	4.8	1.2	Crystal Structure Of The Sam Dependent Methyltransferase From Agrobacterium Tumefaciens
6322393	NP_012467.1	hypothetical protein YJL068C	Saccharomyces cerevisiae S288c	100.0	100.0	100.0	0	1PV1	D Chain D- Crystal Structure Analysis Of Yeast Hypothetical Protein: Yjg8_yeast	3I6Y	267	27.7	10e-28.2	1.3	46.1	2.3	0.5	Structure Of An Esterase From The Oil-Degrading Bacterium Oleispira Antarctica
18977389	NP_578746.1	hypothetical protein PF1017	Pyrococcus furiosus DSM 3638	99.6	99.6	91.2	0	1J31	D Chain D- Crystal Structure Of Hypothetical Protein Ph0642 From Pyrococcus Horikoshii	2VHH	257	27.4	10e-23.7	2	27.2	3.4	0.8	Crystal Structure Of A Pyrimidine Degrading Enzyme From Drosophila Melanogaster
18977663	NP_579020.1	hypothetical protein PF1291	Pyrococcus furiosus DSM 3638	99.6	99.6	92.0	0	2GJU	D Chain D- Crystal Structure Of Hypothetical Protein Ph1004 From Pyrococcus Horikoshii Ot3	3RQZ	211	27.3	10e-23.1	1.9	24.2	NA	NA	Crystal Structure Of Metallophosphoesterase From Sphaerobacter Thermophilus
6319435	NP_009517.1	hypothetical protein YBL036C	Saccharomyces cerevisiae S288c	100.0	100.0	100.0	0	1B54	A Chain A- Crystal Structure Of A Yeast Hypothetical Protein-A Structure From Bnl's Human Proteome Project	3N2O	215	22.5	10e-14.6	3	8.4	NA	NA	X-Ray Crystal Structure Of Arginine Decarboxylase Complexed Arginine From Vibrio Vulnificus
18977487	NP_578844.1	hypothetical protein PF1115	Pyrococcus furiosus DSM 3638	100.0	96.1	95.3	0	3D79	A Chain A- Crystal Structure Of Hypothetical Protein Ph0734.1 From Hyperthermophilic Archaea Pyrococcus Horikoshii Ot3	1ZS7	149	20.9	10e-15.5	2	25.5	2.3	1.5	The Structure Of Gene Product Ape0525 From Aeropyrum Pernix
18977235	NP_578592.1	hypothetical protein PF0863	Pyrococcus furiosus DSM 3638	99.4	95.0	100.0	0	1YEM	B Chain B- Conserved Hypothetical Protein Pfu-838710-001 From Pyrococcus Furiosus	3N10	157	17.5	10e-12.4	2.2	22.3	NA	NA	Product Complex Of Adenylate Cyclase Class Iv
33359473	NP_578084.2	hypothetical protein PF0355	Pyrococcus furiosus DSM 3638	98.7	100.0	93.3	0	1V96	B Chain B- Crystal Structure Of Hypothetical Protein Of Unknown Function From Pyrococcus Horikoshii Ot3	3H87	117	15.7	10e-13.4	2.1	18.8	3.1	2	Rv0301 Rv0300 Toxin Antitoxin Complex From Mycobacterium Tuberculosis
18977279	NP_578636.1	hypothetical protein PF0907	Pyrococcus furiosus DSM 3638	99.1	92.2	98.1	0	1XE1	A Chain A- Hypothetical Protein From Pyrococcus Furiosus Pfu-880080-001	1D1N	81	15.4	10e-12.0	2.3	25.9	2	3.2	Solution Structure Of The Fmet-Trnafmet Binding Domain Of Becillus Stearothermophillus Translation Initiation Factor If2
18977097	NP_578454.1	hypothetical protein PF0725	Pyrococcus furiosus DSM 3638	99.2	93.5	99.2	0	1Y81	A Chain A- Conserved Hypothetical Protein Pfu-723267-001 From Pyrococcus Furiosus	2DUW	112	15.3	10e-14.8	1.7	33	1.2	1.7	Solution Structure Of Putative Coa-Binding Protein Of Klebsiella Pneumoniae

Visual inspection of protein structural alignment could be useful to evaluate protein conformations variation and to study the characteristic of the active sites as illustrated in Figure 4.3.

Figure 4.3a shows the structural alignment between the hypothetical protein PF1291 from *Pyrococcus Furiosus* (RefSeq accession number NP_579020.1) that has an experimentally determined structure (PDB 1NNW) with the structure annotated as "Metallophosphoesterase" from *Sphaerobacter Thermophilus* (PDB 3RQZ). Red tubes indicate identical amino acid in the protein alignment; blue the presence of different type of amino acids; and grey the part of protein structure that was not aligned. This is the case were the hypothetical protein has an experimentally determinate function and there is no need to find similar protein in PDB database.

Figure 4.3b shows the resuls of annotation for hypothetical protein Ph1017 from *Pyrococcus furiosus DSM 3638*. This protein does not have an experimentally determine structure, but it is very similar in amino acid sequence to protein PDB 1J31 from *Pyrococcus Horikoshii*. Its amino acid sequence similarity is greater than 91 percent and the protein coverage for both query and subject protein is greater than 99 percent. The structure of protein PDB 1J31 is higly similar to the structure annotated as "N-Carbamyl-D- Amino Acid Amidohydrolase Complexed With N-Carbamyl-D- Methionine" (PDB 1UF5) as shown in Figure 4.3b. For this reason the hypothetical protein Ph1017 to be investigated could have a function of "N-Carbamyl-D- Amino Acid Amidohydrolase".

Figure 4.3c shows the annotation process for hypothetical protein TK0174 from *Thermococcus kodakarensis KOD1* reported in Table 3.2. This protein does not have a structure in PDB, but it is highly similar to the protein Ph0010 from *Pyrococcus*

Horikoshii that has the structure with PDB identifier equal to 1VAJ. The VAST results for structure 1VAJ provide a hit to the protein described as “Catalytic Core of Dna Polymerase" with PDB identifier equal to 1T3N. In this case the VAST score has a low value of 7.8, and the superposition of the structure showed in 4.3c clearly highlights the fact that the annotation could not be given on these bases. As observed before, structural similarity match must be considered as a valuable tool to protein characterization whenever the classical methods of sequence similarity fail to give a hint of the possible protein function; however the proposed annotation by this method must be carefully evaluated.

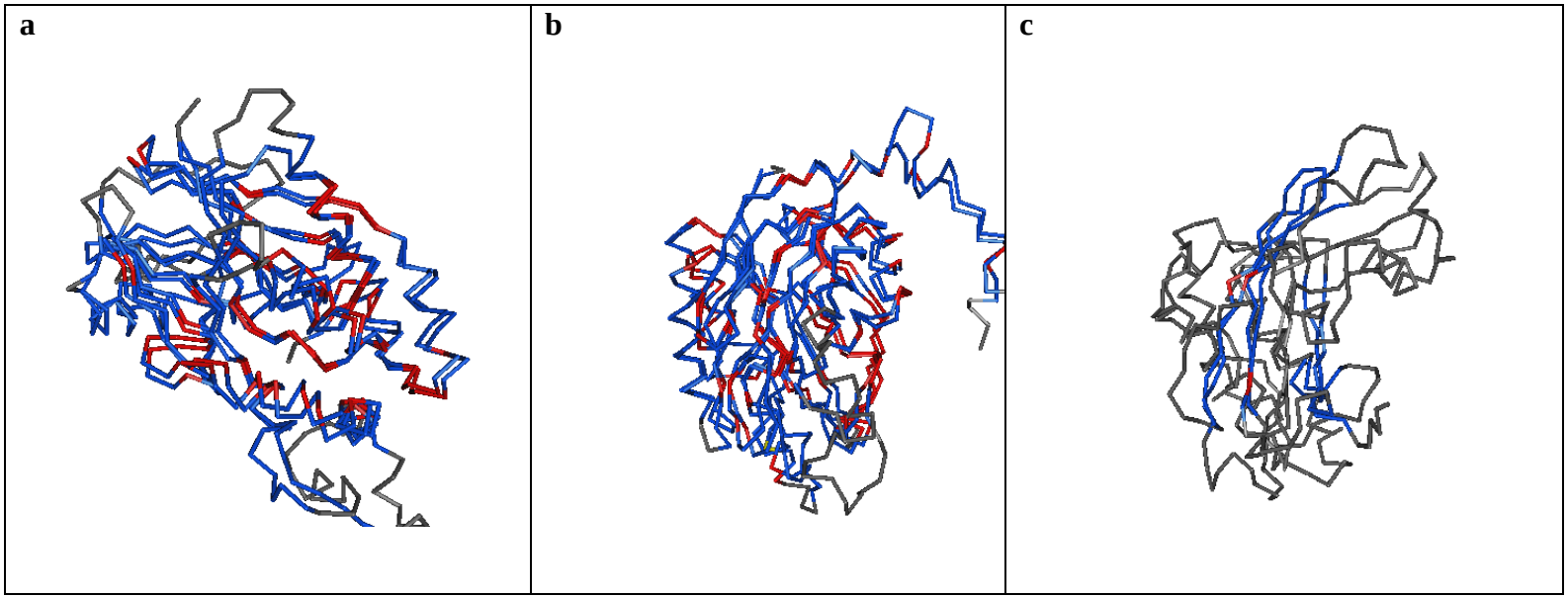


Figure 4.3. VAST visual inspection of structural alignments. A. Structural alignment between the hypothetical protein PF1291 from *Pyrococcus Furiosus* (RefSeq accession number NP_579020.1, PDB 1NNW) with the structure annotated as "Metallophosphoesterase" from *Sphaerobacter Thermophilus* (PDB 3RQZ). B. Structural alignment between hypothetical protein Ph0642 from *Pyrococcus Horikoshii* (PDB 1J31) with the structure annotated as "N-Carbamyl-D- Amino Acid Amidohydrolase Complexed With N-Carbamyl-D- Methionine" (PDB 1UF5). C. Structural alignment between hypothetical protein Ph0010 from *Pyrococcus Horikoshii* (PDB 1VAJ) with the structure annotated as "Catalytic Core of Dna Polymerase" (PDB 1T3N).

CHAPTER 5

CONCLUSION

The software tool, HYPE, was developed to facilitate the annotation of a large number of hypothetical proteins. The performance of the tool was assessed against the proteins of twelve organisms in NCBI RefSeq database. The results were surprising, the function associated to 54 hypothetical proteins in RefSeq could be proposed with the maximum level of accuracy (100% amino acid sequence similarity match and 100% protein coverage). Thousand of hypothetical proteins could be annotated with reduced criteria of similarities. The potential loss of accuracy in protein annotation when criteria of similarities are not very stringent could be mitigated by the introduction of an indication of the quality of annotation. An example of the quality of annotation was proposed. The use of our tool allows performing parametric analysis of protein similarities; this feature could be useful to prioritize hypothetical protein annotation tasks to those proteins with high level of similarities.

Protein function identifications and annotations are dynamic processes; discoveries of cellular functions for a protein in one organism can potentially trigger an update of the annotation of similar proteins in different organisms. Our tool can easily highlight all the proteins affected by the newly proposed functions allowing the propagation of the annotation to other organisms and internal consistency of databases. This is achieved by running recursively the tools for single protein similarity match such as NCBI BLAST and NCBI VAST and combining the results in a unique data file easy to manipulate with ad-hoc developed post-processing tools. The broad use of HYPE for

hypothetical protein annotation could bring a further advantage in the standardization of the annotation text of protein within the database. HYPE search for hypothetical proteins is based on the presence of the keyword “hypothetical” in the annotation text. This could be considered a limiting factor for the retrieval of all the hypothetical proteins in one organism, however it allows a great flexibility in the tool to be used versus other proteins with an undetermined function and annotated in various ways: “uncharacterized”, “unknown function”, “predicted protein” and versus other databases such as GenBank.

HYPE is a modular application that could be extended to incorporate other functionality in order to increase the level of confidence in the protein annotation. It is currently under evaluation the possibility to include the analysis of conserved protein domains maintained in the NCBI Conserved Domain Database (CDD) (Marchler-Bauer, Lu et al. 2011) and the capability to perform multiple sequence alignment searches for phylogenic profiles generation by incorporating tool such as ClustalW2.

REFERENCES

Allen, M. A., F. M. Lauro, et al. (2009). "The genome sequence of the psychrophilic archaeon, Methanococcoides burtonii: the role of genome evolution in cold adaptation." ISME J 3(9): 1012-1035.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res 25(17): 3389-3402.

Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains." Science 181(4096): 223-230.

Atomi, H., T. Fukui, et al. (2004). "Description of Thermococcus kodakaraensis sp. nov., a well studied hyperthermophilic archaeon previously reported as Pyrococcus sp. KOD1." Archaea 1(4): 263-267.

Benson, D. A., I. Karsch-Mizrachi, et al. (2012). "GenBank." Nucleic Acids Res 40(Database issue): D48-53.

Bergman, N. H. (2007). Comparative Genomics, Methods in Molecular Biology, Vol. 395-396.

Bethesda (2008). BLAST Help, National Center for Biotechnology Information.

Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science 277(5331): 1453-1462.

Brunner, A. M., V. B. Busov, et al. (2004). "Poplar genome sequence: functional genomics in an ecologically dominant plant species." Trends Plant Sci 9(1): 49-56.

C.elegans-Sequencing-Consortium (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science 282(5396): 2012-2018.

Cerami, E. (2005). XML for Bioinformatics, Springer Science + BusinessMedia Inc.

Chenna, R., H. Sugawara, et al. (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Res 31(13): 3497-3500.

Gibrat, J. F., T. Madej, et al. (1996). "Surprising similarities in structure comparison." Curr Opin Struct Biol 6(3): 377-385.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A 89(22): 10915-10919.

Kolker, E., K. S. Makarova, et al. (2004). "Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae." Nucleic Acids Res 32(8): 2353-2361.

Kolodny, R., P. Koehl, et al. (2005). "Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures." J Mol Biol 346(4): 1173-1188.

Korf, I., M. Yandell, et al. (2003). BLAST, O'Reilly Media, Inc.

Kour, A., K. Greer, et al. (2011). "MGOS: Development of a Community Annotation Database for Magnaporthe oryzae." Mol Plant Microbe Interact.

Madej, T., K. J. Addess, et al. (2012). "MMDB: 3D structures and macromolecular interactions." Nucleic Acids Res 40(Database issue): D461-464.

Madej, T., J. F. Gibrat, et al. (1995). "Threading a database of protein cores." Proteins 23(3): 356-369.

Marchler-Bauer, A., S. Lu, et al. (2011). "CDD: a Conserved Domain Database for the functional annotation of proteins." Nucleic Acids Res 39(Database issue): D225-229.

McEntyre, J. and J. Ostell (2002). The NCBI Handbook, Nationl Center for Biotechnology Information.

Muzny, D. M., S. E. Scherer, et al. (2006). "The DNA sequence, annotation and analysis of human chromosome 3." Nature 440(7088): 1194-1198.

Panchenko, A. R. and T. Madej (2004). "Analysis of protein homology by assessing the (dis)similarity in protein loop regions." Proteins 57(3): 539-547.

Polyanovsky, V. O., M. A. Roytberg, et al. (2011). "Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences." Algorithms Mol Biol 6(1): 25.

Prakash, A. and M. Tompa (2005). "Statistics of local multiple alignments." Bioinformatics 21 Suppl 1: i344-350.

Pruitt, K. D., T. Tatusova, et al. (2012). "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy." Nucleic Acids Res 40(Database issue): D130-135.

RCSB (2012). RCSB PDB Newletter RCSB Protein Data Bank Newletter, Protein Data Bank.

Robb, F. T., D. L. Maeder, et al. (2001). "Genomic sequence of hyperthermophile, Pyrococcus furiosus: implications for physiology and enzymology." Methods Enzymol 330: 134-157.

Rose, P. W., B. Beran, et al. (2011). "The RCSB Protein Data Bank: redesigned web site and web services." Nucleic Acids Res 39(Database issue): D392-401.

Sager-Institute. (2007). "Mouse Sequencing Consortium." from http://www.sanger.ac.uk/Projects/M_musculus/.

Sanger-Institute. (2001-2012). "Danio rerio sequencing project ", from http://www.sanger.ac.uk/Projects/D_rerio/.

Sayers, E. W., T. Barrett, et al. (2012). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res 40(Database issue): D13-25.

Schwartz, R. and T. Christiansen (1997). Learning Perl, O'Reilly & Associates, Inc.

Shapiro, J. and D. Brutlag (2004). "FoldMiner: structural motif discovery using an improved superposition algorithm." Protein Sci 13(1): 278-294.

Sierk, M. L. and W. R. Pearson (2004). "Sensitivity and selectivity in protein structure comparison." Protein Sci 13(3): 773-785.

Sivashankari, S. and P. Shanmughavel (2006). "Functional annotation of hypothetical proteins - A review." Bioinformation 1(8): 335-338.

The-Arabidopsis-Genome-Initiative (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature 408(6814): 796-815.

Wood, V., K. M. Rutherford, et al. (2001). "A re-annotation of the Saccharomyces cerevisiae genome." Comp Funct Genomics 2(3): 143-154.

Wrabl, J. O. and N. V. Grishin (2008). "Statistics of random protein superpositions: p-values for pairwise structure alignment." J Comput Biol 15(3): 317-355.

Zarembinski, T. I., L. W. Hung, et al. (1998). "Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics." Proc Natl Acad Sci U S A 95(26): 15189-15193.

APPENDIX A

Table A.1. HYPE proposed annotations that satisfy the criteria of amino acid sequence having similarities greater than 98% and hypothetical protein coverage greater than 98%.

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
57639990	YP_182468.1	hypothetical protein TK0055	Thermococcus kodakarensis KOD1	100.0	100.0	98.2	0.0	ZP_04878637.1	ProFAR isomerase associated superfamily protein	Thermococcus sp. AM4
57640293	YP_182771.1	hypothetical protein TK0358	Thermococcus kodakarensis KOD1	99.4	100.0	98.5	0.0	YP_002307099.1	RNA terminal phosphate cyclase	Thermococcus onnurineus NA1
57641158	YP_183636.1	hypothetical protein TK1223	Thermococcus kodakarensis KOD1	100.0	100.0	98.9	0.0	ZP_04878882.1	membrane bound hydrogenase- MbxD subunit	Thermococcus sp. AM4
57641213	YP_183691.1	hypothetical protein TK1278	Thermococcus kodakarensis KOD1	100.0	100.0	99.1	0.0	ZP_04880486.1	DNA-binding protein	Thermococcus sp. AM4
57641496	YP_183974.1	hypothetical protein TK1561	Thermococcus kodakarensis KOD1	99.0	98.5	98.0	0.0	YP_002958671.1	Multiple antibiotic resistance (Mar)-related protein	Thermococcus gammatolerans EJ3
57641525	YP_184003.1	hypothetical protein TK1590	Thermococcus kodakarensis KOD1	100.0	100.0	98.8	0.0	ZP_04879917.1	RecA-superfamily ATPase implicated in signal transduction	Thermococcus sp. AM4
91772521	YP_565213.1	hypothetical protein Mbur_0479	Methanococcoides burtonii DSM 6242	100.0	100.0	99.2	0.0	YP_565266.1	transposase	Methanococcoides burtonii DSM 6242
91772521	YP_565213.1	hypothetical protein Mbur_0479	Methanococcoides burtonii DSM 6242	100.0	100.0	99.2	0.0	YP_565143.1	transposase	Methanococcoides burtonii DSM 6242
94541112	YP_588454.1	hypothetical protein- Qin prophage	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	98.7	0.0	YP_002411373.1	ynfO qin prophage	Escherichia coli FVEC1302
145698240	YP_001165313.1	hypothetical protein b4592	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	100.0	0.0	ZP_08353023.1	putative periplasmic protein	Escherichia coli M718
145698240	YP_001165313.1	hypothetical protein b4592	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	100.0	0.0	ZP_08358008.1	putative periplasmic protein	Escherichia coli TA206
145698240	YP_001165313.1	hypothetical protein b4592	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	100.0	0.0	ZP_08373304.1	putative periplasmic protein	Escherichia coli TA280

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
145698270	YP_001165321.1	hypothetical protein b4602	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	100.0	0.0	YP_669509.1	membrane protein	Escherichia coli H736
145698284	YP_001165322.1	hypothetical protein b4604	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	100.0	0.0	ZP_03003264.1	uncharacterized protein yojO	Escherichia coli 1827-70
145698342	YP_001165334.1	hypothetical protein b4622	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	98.9	0.0	ZP_06660209.1	ytcA protein	Escherichia coli B185
145698342	YP_001165334.1	hypothetical protein b4622	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	98.9	0.0	ZP_06988134.1	ytcA protein	Escherichia coli FVEC1302
145698342	YP_001165334.1	hypothetical protein b4622	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	100.0	0.0	YP_001746475.1	formate dehydrogenase H	Escherichia coli TA280
145698342	YP_001165334.1	hypothetical protein b4622	Escherichia coli str. K-12 substr. MG1655	100.0	100.0	98.9	0.0	ZP_03030554.1	formate dehydrogenase H	Escherichia coli W
39946336	XP_362705.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.4	0.0	XP_958786.1	proteasome component Y13	Neurospora crassa OR74A
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	98.5	0.0	XP_001537373.1	26S protease regulatory subunit 8	Ajellomyces capsulatus NAm1
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	98.2	0.0	XP_002620460.1	26S protease regulatory subunit 8	Ajellomyces dermatitidis SLH14081
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	99.0	0.0	XP_002849889.1	26S protease regulatory subunit 8	Arthroderma otae CBS 113480
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	99.5	0.0	XP_001547053.1	26S protease regulatory subunit 8	Botryotinia fuckeliana B05.10
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	98.7	0.0	XP_001241366.1	26S proteasome subunit P45 family protein	Coccidioides posadasii C735 delta SOWgp
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	98.5	0.0	XP_003049600.1	26S proteasome regulatory complex- ATPase RPT6	Nectria haematococca mpVI 77-13-4
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	99.5	0.0	XP_963354.1	26S protease regulatory subunit 8	Neurospora crassa OR74A

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	98.5	0.0	XP_002796471.1	26S protease regulatory subunit 8	Paracoccidioides brasiliensis Pb01
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	99.2	0.0	XP_002148124.1	proteasome regulatory particle subunit Rpt6-putative	Penicillium marneffei ATCC 18224
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	99.2	0.0	XP_002482365.1	proteasome regulatory particle subunit Rpt6-putative	Talaromyces stipitatus ATCC 10500
39951475	XP_363454.1	hypothetical protein MGG_01380	Magnaporthe oryzae 70-15	100.0	100.0	99.0	0.0	XP_003230954.1	26S protease regulatory subunit 8	Trichophyton rubrum CBS 118892
39973575	XP_368178.1	hypothetical protein MGG_01066	Magnaporthe oryzae 70-15	100.0	100.0	99.3	0.0	XP_001557680.1	40S ribosomal protein S23	Botryotinia fuckeliana B05.10
39973575	XP_368178.1	hypothetical protein MGG_01066	Magnaporthe oryzae 70-15	100.0	100.0	98.6	0.0	XP_002144194.1	ribosomal protein S23 (S12)	Penicillium marneffei ATCC 18224
39973575	XP_368178.1	hypothetical protein MGG_01066	Magnaporthe oryzae 70-15	100.0	100.0	99.3	0.0	XP_001588322.1	40S ribosomal protein S23	Sclerotinia sclerotiorum 1980
39973575	XP_368178.1	hypothetical protein MGG_01066	Magnaporthe oryzae 70-15	100.0	100.0	98.6	0.0	XP_002341358.1	ribosomal protein S23 (S12)	Talaromyces stipitatus ATCC 10500
39974775	XP_368778.1	hypothetical protein MGG_00466	Magnaporthe oryzae 70-15	100.0	100.0	98.5	1.0	XP_680756.1	CD42_CHICK Cell division control protein 42 homolog (G25K GTP-binding protein)	Aspergillus nidulans FGSC A4
39974775	XP_368778.1	hypothetical protein MGG_00466	Magnaporthe oryzae 70-15	99.5	99.0	98.4	0.0	XP_385623.1	CD42_CHICK Cell division control protein 42 homolog (G25K GTP-binding protein)	Gibberella zeae PH-1
149210305	XP_001522527.1	hypothetical protein MGCH7_ch7g630	Magnaporthe oryzae 70-15	100.0	100.0	100.0	0.0	XP_965758.1	40S ribosomal protein S27	Neurospora crassa OR74A
149210305	XP_001522527.1	hypothetical protein MGCH7_ch7g630	Magnaporthe oryzae 70-15	100.0	100.0	100.0	0.0	XP_366796.1	40S ribosomal protein S27	Verticillium albo-atrum VaMs.102
39945474	XP_362274.1	hypothetical protein MGG_04719	Magnaporthe oryzae 70-15	100.0	100.0	98.7	0.0	XP_001226350.1	guanine nucleotide-binding protein beta subunit-like protein	Chaetomium globosum CBS 148.51

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
39945474	XP_362274.1	hypothetical protein MGG_04719	Magnaporthe oryzae 70-15	100.0	100.0	98.4	0.0	XP_390046.1	GBLP_NEUCR Guanine nucleotide-binding protein beta subunit-like protein (Cross-pathway control WD-repeat protein cpc-2)	Gibberella zeae PH-1
39945474	XP_362274.1	hypothetical protein MGG_04719	Magnaporthe oryzae 70-15	100.0	100.0	98.7	0.0	XP_003002475.1	G-protein beta subunit	Verticillium albo-atrum VaMs.102
39951877	XP_363655.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.2	0.0	XP_001558921.1	26S protease regulatory subunit 7	Botryotinia fuckeliana B05.10
39951877	XP_363655.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.9	0.0	XP_964130.1	26S protease regulatory subunit 7	Neurospora crassa OR74A
39951877	XP_363655.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	99.1	0.0	XP_003005299.1	26S protease regulatory subunit 7	Verticillium albo-atrum VaMs.102
39952239	XP_363836.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	99.3	98.6	0.0	XP_001558601.1	clathrin coat assembly protein	Botryotinia fuckeliana B05.10
39952239	XP_363836.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.6	0.0	XP_962659.1	AP-2 complex subunit sigma	Neurospora crassa OR74A
39952239	XP_363836.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.6	0.0	XP_003009348.1	AP-2 complex subunit sigma	Verticillium albo-atrum VaMs.102
39973891	XP_368336.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.1	0.0	XP_003005676.1	26S protease regulatory subunit 6B	Verticillium albo-atrum VaMs.102
39976861	XP_369818.1	hypothetical protein MGG_06333	Magnaporthe oryzae 70-15	100.0	100.0	98.5	0.0	XP_001229506.1	60S ribosomal protein L29	Chaetomium globosum CBS 148.51
145603314	XP_001404467.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.6	1.4	XP_001728202.1	DASH complex subunit DAD4	Neurospora crassa OR74A
145603314	XP_001404467.1	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.6	1.4	XP_002152784.1	DASH complex subunit Dad4- putative	Penicillium marneffei ATCC 18224
145608394	XP_360608.2	conserved hypothetical protein	Magnaporthe oryzae 70-15	100.0	100.0	98.2	0.0	XP_956516.2	maintenance of ploidy protein mob1	Neurospora crassa OR74A
145616510	XP_366415.2	conserved hypothetical protein	Magnaporthe oryzae 70-15	99.5	99.5	98.4	0.5	XP_001259722.1	AP-1 adaptor complex subunit mu- putative	Neosartorya fischeri NRRL 181
145603147	XP_361955.2	hypothetical protein MGG_04400	Magnaporthe oryzae 70-15	99.7	99.2	98.2	0.0	XP_001561381.1	eukaryotic initiation factor 4A	Botryotinia fuckeliana B05.10
145603147	XP_361955.2	hypothetical protein MGG_04400	Magnaporthe oryzae 70-15	99.5	99.2	98.2	0.0	XP_001220184.1	cell cycle control protein-related	Chaetomium globosum CBS 148.51

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
145603147	XP_361955.2	hypothetical protein MGG_04400	Magnaporthe oryzae 70-15	99.5	99.2	99.0	0.0	XP_958421.2	eIF4A	Neurospora crassa OR74A
145603147	XP_361955.2	hypothetical protein MGG_04400	Magnaporthe oryzae 70-15	99.7	99.2	98.2	0.0	XP_001594651.1	eukaryotic initiation factor 4A	Sclerotinia sclerotiorum 1980
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	98.7	0.0	XP_001542533.1	40S ribosomal protein S15	Ajellomyces capsulatus NAm1
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	98.7	0.0	XP_002625565.1	40S ribosomal protein S15	Ajellomyces dermatitidis SLH14081
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	98.7	0.0	XP_002843919.1	40S ribosomal protein S15	Arthroderma otae CBS 113480
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	98.0	0.0	XP_001553822.1	40S ribosomal protein S15	Botryotinia fuckeliana B05.10
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	98.0	0.0	XP_001248452.1	40S ribosomal protein S15- putative	Coccidioides posadasii C735 delta SOWgp
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	99.3	0.7	XP_380571.1	RS15_PODAN 40S RIBOSOMAL PROTEIN S15 (S12)	Gibberella zeae PH-1
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	98.7	0.0	XP_002793160.1	40S ribosomal protein S15	Paracoccidioides brasiliensis Pb01
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	98.0	0.0	XP_002565926.1	Pc22g20260	Penicillium chrysogenum Wisconsin 54-1255
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	98.0	0.0	XP_001597504.1	40S ribosomal protein S15	Sclerotinia sclerotiorum 1980
145614766	XP_366150.2	hypothetical protein MGG_10370	Magnaporthe oryzae 70-15	100.0	100.0	99.3	0.0	XP_003015339.1	40S ribosomal protein S15	Trichophyton rubrum CBS 118892
149210177	XP_001522463.1	hypothetical protein MG02813.4	Magnaporthe oryzae 70-15	100.0	100.0	99.8	0.2	XP_366737.2	malate synthase A	Magnaporthe oryzae 70-15
13449319	NP_085501.1	hypothetical protein ArthMp033	Arabidopsis thaliana	100.0	100.0	99.5	0.0	NP_178809.1	Ycf1 protein	Arabidopsis thaliana
13449380	NP_085562.1	hypothetical protein ArthMp095	Arabidopsis thaliana	100.0	100.0	99.2	0.0	NP_973438.1	ATP synthase 9	Arabidopsis thaliana

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
13449398	NP_085580.1	hypothetical protein ArthMp104	Arabidopsis thaliana	100.0	100.0	99.7	0.0	NP_178786.1	cytochrome c oxidase subunit II	Arabidopsis thaliana
41053459	NP_956606.1	hypothetical protein LOC393282	Danio rerio	100.0	100.0	99.2	0.0	NP_955829.1	NHP2 non-histone chromosome protein 2-like 1	Danio rerio
41053459	NP_956606.1	hypothetical protein LOC393282	Danio rerio	100.0	100.0	99.2	0.0	NP_001187959.1	nhp2-like protein 1	Ictalurus punctatus
41053911	NP_956269.1	hypothetical protein LOC335798	Danio rerio	99.8	99.3	98.2	0.2	XP_001656025.1	tubulin beta chain	Aedes aegypti
41053911	NP_956269.1	hypothetical protein LOC335798	Danio rerio	99.3	99.5	98.4	0.2	NP_509585.1	C. briggsae CBR-TBB-4 protein	Caenorhabditis briggsae
41053911	NP_956269.1	hypothetical protein LOC335798	Danio rerio	98.2	100.0	98.2	0.0	XP_002579105.1	tubulin subunit beta	Schistosoma mansoni
41152297	NP_957010.1	hypothetical protein LOC393689	Danio rerio	100.0	100.0	98.1	0.0	NP_001187577.1	upf0139 membrane protein c19orf56-like protein	Ictalurus punctatus
41152382	NP_956303.1	hypothetical protein LOC336334	Danio rerio	100.0	100.0	99.8	0.0	NP_001017795.1	eukaryotic translation elongation factor 1 alpha 1	Danio rerio
41152382	NP_956303.1	hypothetical protein LOC336334	Danio rerio	100.0	100.0	98.7	0.0	NP_001167438.1	Elongation factor 1-alpha 1	Salmo salar
47086529	NP_997925.1	hypothetical protein LOC336641	Danio rerio	100.0	100.0	99.5	0.0	NP_001187044.1	ribosomal protein L17	Ictalurus punctatus
50344940	NP_001002142.1	hypothetical protein LOC415232	Danio rerio	100.0	100.0	99.4	0.0	NP_001002582.1	crystallin- gamma M2d15	Danio rerio
50344940	NP_001002142.1	hypothetical protein LOC415232	Danio rerio	100.0	100.0	98.3	0.0	NP_001038331.1	crystallin- gamma M2d2	Danio rerio
50540458	NP_001002693.1	hypothetical protein LOC436966	Danio rerio	100.0	100.0	100.0	0.0	NP_001104721.1	receptor activity-modifying protein 1	Danio rerio
51011055	NP_001003485.1	hypothetical protein LOC445091	Danio rerio	100.0	100.0	98.1	0.0	NP_001017807.1	calpain-2 catalytic subunit	Danio rerio
54400556	NP_001006027.1	hypothetical protein LOC450006	Danio rerio	98.9	99.4	98.3	0.0	NP_998377.1	myosin- light polypeptide 9- regulatory	Danio rerio
56090176	NP_001007770.1	hypothetical protein LOC493609	Danio rerio	100.0	100.0	98.2	0.0	NP_998200.1	growth factor receptor-bound protein 2	Danio rerio
56090176	NP_001007770.1	hypothetical protein LOC493609	Danio rerio	100.0	100.0	98.2	0.0	NP_001187313.1	growth factor receptor-bound protein 2	Ictalurus punctatus
61806482	NP_001013473.1	hypothetical protein LOC541327	Danio rerio	100.0	100.0	98.5	0.0	NP_001139966.1	40S ribosomal protein S17	Ictalurus punctatus
62955177	NP_001017600.1	hypothetical protein LOC550263	Danio rerio	100.0	100.0	99.3	0.0	NP_001140124.1	DNA-directed RNA polymerases I- II- and III subunit RPABC3	Salmo salar

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
62955177	NP_001017600.1	hypothetical protein LOC550263	Danio rerio	100.0	100.0	98.7	0.0	NP_001016113.1	polymerase (RNA) II (DNA directed) polypeptide H	Xenopus (Silurana) tropicalis
62955177	NP_001017600.1	hypothetical protein LOC550263	Danio rerio	100.0	100.0	98.7	0.0	NP_001085470.1	polymerase (RNA) II (DNA directed) polypeptide H	Xenopus laevis
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	99.3	100.0	99.3	0.0	XP_001657153.1	histone H3	Aedes aegypti
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	99.3	100.0	98.5	0.0	XP_001655332.1	histone H3	Aedes aegypti
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	99.3	100.0	99.3	0.0	XP_307081.3	Anopheles gambiae str. PEST AGAP012709-PA	Anopheles gambiae str. PEST
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_320335.2	AGAP012196-PA	Anopheles gambiae str. PEST
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	NP_189372.1	histone H3	Arabidopsis lyrata subsp. lyrata
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	99.3	0.0	XP_001892001.1	histone H3	Brugia malayi
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	NP_496899.1	HIStone family member (his-2)	Caenorhabditis elegans
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	99.3	99.3	98.5	0.0	XP_001853968.1	histone H3.2	Culex quinquefasciatus
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002045560.1	GM26661	Drosophila sechellia
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002045486.1	GM19351	Drosophila sechellia
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002045405.1	GM13629	Drosophila sechellia
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002045602.1	GM25125	Drosophila sechellia
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002045660.1	GM18832	Drosophila sechellia
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002045096.1	GM19740	Drosophila sechellia
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002045506.1	GM11653	Drosophila sechellia
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_001967747.1	GK25031	Drosophila willistoni
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002075944.1	GK12402	Drosophila willistoni

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	99.3	0.0	XP_002502604.1	histone H3	Micromonas pusilla CCMP1545
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_001083065.1	histone H3- putative	Perkinsus marinus ATCC 50983
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002775920.1	histone H3- putative	Perkinsus marinus ATCC 50983
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	99.3	0.0	XP_001753449.1	histone H3	Physcomitrella patens subsp. patens
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002999294.1	histone H3	Phytophthora infestans T30-4
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002578635.1	histone H3	Schistosoma mansoni
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002579313.1	histone H3	Schistosoma mansoni
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	NP_999712.1	histone H3- embryonic	Strongylocentrotus purpuratus
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002177514.1	histone H3	Thalassiosira pseudonana CCMP1335
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	XP_002365267.1	histone H3	Toxoplasma gondii ME49
68448483	NP_001020342.1	hypothetical protein LOC573992	Danio rerio	100.0	100.0	98.5	0.0	NP_001091119.1	histone cluster 1- H3g protein	Xenopus laevis
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.7	0.2	XP_001655855.1	tubulin alpha chain	Aedes aegypti
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.7	0.0	NP_001036885.1	alpha-tubulin	Bombyx mori
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.7	0.2	NP_001029376.1	tubulin alpha-1C chain	Bos taurus
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.9	0.2	XP_001861873.1	tubulin alpha-2 chain	Culex quinquefasciatus
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.9	0.2	NP_919369.2	tubulin alpha-1C chain	Danio rerio
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.4	0.2	XP_002073900.1	GK14362	Drosophila willistoni
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.7	0.2	NP_116093.1	tubulin alpha-1C chain	Homo sapiens
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.7	0.2	XP_002416661.1	alpha tubulin	Ixodes scapularis

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.7	0.2	XP_002402152.1	alpha tubulin	Ixodes scapularis
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	99.8	99.1	99.1	0.2	XP_002427228.1	tubulin alpha-3 chain-putative	Pediculus humanus corporis
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.7	0.2	NP_001011995.1	tubulin alpha-1C chain	Rattus norvegicus
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.4	0.2	NP_001015934.1	tubulin- alpha 3c	Xenopus (Silurana) tropicalis
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.4	0.2	NP_001165669.1	tubulin- alpha 3e	Xenopus laevis
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.4	0.2	NP_001095253.1	tubulin alpha chain	Xenopus laevis
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.4	0.2	NP_989078.1	tubulin- alpha 6	Xenopus laevis
83025060	NP_001032649.1	hypothetical protein LOC641561	Danio rerio	100.0	100.0	98.2	0.0	NP_001079523.1	tubulin- alpha 3c	Xenopus laevis
89886305	NP_001034907.1	hypothetical protein LOC562838	Danio rerio	100.0	100.0	98.4	0.0	NP_956065.1	ras-related C3 botulinum toxin substrate 1	Danio rerio
89886305	NP_001034907.1	hypothetical protein LOC562838	Danio rerio	100.0	100.0	98.4	0.0	NP_001089332.1	ras-related C3 botulinum toxin substrate 1 (rho family- small GTP binding protein Rac1)	Xenopus laevis
89886305	NP_001034907.1	hypothetical protein LOC562838	Danio rerio	100.0	100.0	98.4	0.0	NP_001084224.1	ras-related C3 botulinum toxin substrate 3 (rho family- small GTP binding protein Rac3)	Xenopus laevis
113675575	NP_001038702.1	hypothetical protein LOC692260	Danio rerio	100.0	100.0	99.5	0.0	NP_571052.1	zinc finger protein draculin	Danio rerio
113677685	NP_001038355.1	hypothetical protein LOC559296	Danio rerio	100.0	100.0	98.4	0.0	NP_001076414.2	guanylate binding protein 4	Danio rerio
113680623	NP_001038923.1	hypothetical protein LOC751748	Danio rerio	100.0	100.0	98.8	0.0	NP_001076353.1	trace amine associated receptor 14b	Danio rerio
113680623	NP_001038923.1	hypothetical protein LOC751748	Danio rerio	100.0	100.0	98.5	0.0	NP_001076381.1	trace amine associated receptor 14d	Danio rerio
115496922	NP_001070117.1	hypothetical protein LOC767711	Danio rerio	100.0	100.0	100.0	0.0	NP_001007786.1	crystallin- gamma M1	Danio rerio
115529301	NP_001070183.1	hypothetical protein LOC767746	Danio rerio	99.8	99.3	98.2	0.2	XP_001656025.1	tubulin beta chain	Aedes aegypti
115529301	NP_001070183.1	hypothetical protein LOC767746	Danio rerio	99.8	99.3	98.2	0.2	XP_309765.4	AGAP010929-PA	Anopheles gambiae str. PEST

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
115529301	NP_001070183.1	hypothetical protein LOC767746	Danio rerio	99.3	99.5	98.2	0.2	NP_509585.1	C. briggsae CBR-TBB-4 protein	Caenorhabditis briggsae
115529379	NP_001070217.1	hypothetical protein LOC767782	Danio rerio	100.0	100.0	99.0	0.0	NP_001093194.1	histone H4-like	Bos taurus
115529379	NP_001070217.1	hypothetical protein LOC767782	Danio rerio	99.0	98.1	100.0	0.0	XP_001865498.1	Histone H4	Culex quinquefasciatus
115529379	NP_001070217.1	hypothetical protein LOC767782	Danio rerio	100.0	100.0	99.0	0.0	NP_001099176.1	histone 1- H4- like	Danio rerio
115529379	NP_001070217.1	hypothetical protein LOC767782	Danio rerio	100.0	100.0	99.0	0.0	NP_001070058.1	histone cluster 1- H4-like	Danio rerio
115529379	NP_001070217.1	hypothetical protein LOC767782	Danio rerio	100.0	100.0	98.1	0.0	XP_002045915.1	GM13181	Drosophila sechellia
115529379	NP_001070217.1	hypothetical protein LOC767782	Danio rerio	100.0	100.0	99.0	0.0	NP_492641.1	late histone L2 H4	Loa loa
116812893	NP_001019569.2	hypothetical protein LOC554097	Danio rerio	100.0	100.0	98.4	0.0	NP_001122231.1	histone 2- H2- like	Danio rerio
116812893	NP_001019569.2	hypothetical protein LOC554097	Danio rerio	100.0	100.0	98.4	0.0	NP_001103306.1	histone H2B	Xenopus (Silurana) tropicalis
121583679	NP_001073537.1	iduronate 2-sulfatase	Danio rerio	100.0	100.0	99.5	0.0	NP_001074043.1	iduronate 2-sulfatase	Danio rerio
125819795	XP_001334176.1	PREDICTED: hypothetical protein LOC797638 isoform 1	Danio rerio	100.0	100.0	98.8	0.0	NP_001107105.1	zinc finger-like gene 1	Danio rerio
125819795	XP_001334176.1	PREDICTED: hypothetical protein LOC797638 isoform 1	Danio rerio	100.0	100.0	98.0	0.0	NP_001139173.1	similar to zinc finger-like gene 1	Danio rerio
125819795	XP_001334176.1	PREDICTED: hypothetical protein LOC797638 isoform 1	Danio rerio	100.0	100.0	98.0	0.0	NP_001164503.1	zinc finger-like gene 1	Danio rerio
125825258	XP_001337285.1	PREDICTED: hypothetical protein LOC799877 isoform 2	Danio rerio	100.0	100.0	99.2	0.0	NP_001107105.1	zinc finger-like gene 1	Danio rerio
125825258	XP_001337285.1	PREDICTED: hypothetical protein LOC799877 isoform 2	Danio rerio	100.0	100.0	98.0	0.0	NP_001139173.1	similar to zinc finger-like gene 1	Danio rerio

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
125845534	XP_001336596.1	PREDICTED: hypothetical protein LOC796282	Danio rerio	100.0	100.0	98.7	0.0	NP_001093875.1	naked cuticle homolog 3	Danio rerio
125845538	XP_001336743.1	PREDICTED: hypothetical protein LOC796413	Danio rerio	100.0	100.0	99.1	0.0	NP_001093875.1	naked cuticle homolog 3	Danio rerio
125846550	XP_001335031.1	PREDICTED: hypothetical protein LOC797777	Danio rerio	100.0	100.0	98.2	0.0	NP_001107105.1	zinc finger-like gene 1	Danio rerio
125847245	XP_001332345.1	PREDICTED: hypothetical protein LOC793437	Danio rerio	100.0	100.0	98.6	0.0	NP_001107105.1	zinc finger-like gene 1	Danio rerio
125847245	XP_001332345.1	PREDICTED: hypothetical protein LOC793437	Danio rerio	100.0	100.0	98.2	0.0	NP_001164503.1	zinc finger-like gene 1	Danio rerio
125851898	XP_001335379.1	PREDICTED: hypothetical protein LOC796596	Danio rerio	100.0	100.0	100.0	0.0	NP_001107105.1	zinc finger-like gene 1	Danio rerio
125851898	XP_001335379.1	PREDICTED: hypothetical protein LOC796596	Danio rerio	100.0	100.0	98.0	0.0	NP_001139173.1	similar to zinc finger-like gene 1	Danio rerio
130491697	NP_001076302.1	hypothetical protein LOC562963	Danio rerio	100.0	100.0	98.5	0.0	NP_001098600.1	mitochondrial inner membrane protein OXA1L	Danio rerio
148229711	NP_001083038.1	hypothetical protein LOC100038789	Danio rerio	100.0	100.0	98.1	0.0	NP_001093194.1	histone H4-like	Bos taurus
148229711	NP_001083038.1	hypothetical protein LOC100038789	Danio rerio	99.0	98.1	99.0	0.0	XP_001865498.1	Histone H4	Culex quinquefasciatus
148229711	NP_001083038.1	hypothetical protein LOC100038789	Danio rerio	100.0	100.0	98.1	0.0	NP_001099176.1	histone 1- H4- like	Danio rerio
148229711	NP_001083038.1	hypothetical protein LOC100038789	Danio rerio	100.0	100.0	98.1	0.0	NP_001070058.1	histone cluster 1- H4-like	Danio rerio
148229711	NP_001083038.1	hypothetical protein LOC100038789	Danio rerio	100.0	100.0	98.1	0.0	NP_492641.1	late histone L2 H4	Loa loa
149773528	NP_001092713.1	hypothetical protein LOC561929	Danio rerio	100.0	100.0	99.4	0.0	NP_001038572.1	crystallin- gamma M2d7	Danio rerio
149773528	NP_001092713.1	hypothetical protein LOC561929	Danio rerio	100.0	100.0	99.4	0.0	NP_001073530.1	crystallin- gamma M2d4	Danio rerio

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
149773528	NP_001092713.1	hypothetical protein LOC561929	Danio rerio	99.4	98.9	98.8	0.0	NP_001002138.1	crystallin- gamma M2d6	Danio rerio
149773528	NP_001092713.1	hypothetical protein LOC561929	Danio rerio	100.0	100.0	98.3	0.0	NP_001038328.1	crystallin- gamma M2d5	Danio rerio
149773528	NP_001092713.1	hypothetical protein LOC561929	Danio rerio	100.0	100.0	98.3	0.0	NP_001076408.1	crystallin- gamma family-like	Danio rerio
149773528	NP_001092713.1	hypothetical protein LOC561929	Danio rerio	99.4	98.9	98.8	0.0	NP_001018517.1	crystallin- gamma M2d1	Danio rerio
156139167	NP_001095864.1	hypothetical protein LOC797707	Danio rerio	100.0	100.0	98.3	0.0	NP_001002581.2	crystallin- gamma M2d8	Danio rerio
156616354	NP_001096101.1	hypothetical protein LOC100124604	Danio rerio	99.4	100.0	99.4	0.0	NP_001002582.1	crystallin- gamma M2d15	Danio rerio
156616354	NP_001096101.1	hypothetical protein LOC100124604	Danio rerio	99.4	100.0	98.3	0.0	NP_001038331.1	crystallin- gamma M2d2	Danio rerio
156616356	NP_001096102.1	hypothetical protein LOC100124605	Danio rerio	100.0	100.0	99.2	0.0	NP_001103305.1	Zgc:171937 protein	Danio rerio
156616356	NP_001096102.1	hypothetical protein LOC100124605	Danio rerio	100.0	100.0	98.4	0.0	NP_001103306.1	histone H2B	Xenopus (Silurana) tropicalis
156616380	NP_001096114.1	hypothetical protein LOC100124618	Danio rerio	100.0	100.0	98.3	0.0	NP_001002581.2	crystallin- gamma M2d8	Danio rerio
156616380	NP_001096114.1	hypothetical protein LOC100124618	Danio rerio	100.0	100.0	98.3	0.0	NP_001096101.1	Zgc:173493 protein	Danio rerio
156739275	NP_001096585.1	hypothetical protein LOC563390	Danio rerio	100.0	100.0	98.2	0.3	NP_571273.1	cathepsin L- 1 b	Danio rerio
156739281	NP_001096588.1	hypothetical protein LOC564906	Danio rerio	100.0	100.0	98.2	0.3	NP_571273.1	cathepsin L- 1 b	Danio rerio
156739307	NP_001096601.1	hypothetical protein LOC795883	Danio rerio	100.0	100.0	99.7	0.0	NP_001013322.1	natterin-like protein	Danio rerio
157311713	NP_001098585.1	hypothetical protein LOC564979	Danio rerio	100.0	100.0	98.5	0.0	NP_571273.1	cathepsin L- 1 b	Danio rerio
157787177	NP_001099150.1	hypothetical protein LOC564835	Danio rerio	100.0	100.0	98.2	0.0	NP_571273.1	cathepsin L- 1 b	Danio rerio
157954458	NP_001103305.1	hypothetical protein LOC100126106	Danio rerio	100.0	100.0	98.4	0.0	NP_001122231.1	histone 2- H2- like	Danio rerio

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
157954458	NP_001103305.1	hypothetical protein LOC100126106	Danio rerio	100.0	100.0	98.4	0.0	NP_001103306.1	histone H2B	Xenopus (Silurana) tropicalis
158262065	NP_001103411.1	hypothetical protein LOC792049	Danio rerio	100.0	100.0	98.4	0.0	NP_001007389.1	tetratricopeptide repeat protein 36	Danio rerio
160333322	NP_001103752.1	hypothetical protein LOC792137	Danio rerio	100.0	100.0	100.0	0.0	NP_919360.1	carbonyl reductase [NADPH] 1	Danio rerio
162138968	NP_001104662.1	hypothetical protein LOC567623	Danio rerio	100.0	100.0	98.2	0.0	NP_571273.1	cathepsin L- 1 b	Danio rerio
165972449	NP_001107098.1	hypothetical protein LOC792506	Danio rerio	100.0	100.0	99.6	0.0	NP_001170806.1	UDP glucuronosyltransferase 2 family- polypeptide B6	Danio rerio
167555152	NP_001107928.1	hypothetical protein LOC797209	Danio rerio	100.0	100.0	98.6	0.0	NP_956440.1	DNA-(apurinic or apyrimidinic site) lyase 2	Danio rerio
168823524	NP_001108394.1	hypothetical protein LOC100141357	Danio rerio	99.4	98.9	98.8	0.0	NP_001038331.1	crystallin- gamma M2d2	Danio rerio
168823524	NP_001108394.1	hypothetical protein LOC100141357	Danio rerio	100.0	100.0	98.3	0.0	NP_001073530.1	crystallin- gamma M2d4	Danio rerio
168823524	NP_001108394.1	hypothetical protein LOC100141357	Danio rerio	100.0	100.0	98.3	0.0	NP_001038572.1	crystallin- gamma M2d7	Danio rerio
181330164	NP_001116765.1	hypothetical protein LOC569000	Danio rerio	100.0	100.0	99.4	0.0	NP_001038572.1	crystallin- gamma M2d7	Danio rerio
181330164	NP_001116765.1	hypothetical protein LOC569000	Danio rerio	100.0	100.0	99.4	0.0	NP_001073530.1	crystallin- gamma M2d4	Danio rerio
181330164	NP_001116765.1	hypothetical protein LOC569000	Danio rerio	100.0	100.0	98.3	0.0	NP_001038328.1	crystallin- gamma M2d5	Danio rerio
181330164	NP_001116765.1	hypothetical protein LOC569000	Danio rerio	100.0	100.0	98.3	0.0	NP_001076408.1	crystallin- gamma family-like	Danio rerio
181330164	NP_001116765.1	hypothetical protein LOC569000	Danio rerio	99.4	98.9	98.8	0.0	NP_001018517.1	crystallin- gamma M2d1	Danio rerio
181330164	NP_001116765.1	hypothetical protein LOC569000	Danio rerio	99.4	98.9	98.8	0.0	NP_001002138.1	crystallin- gamma M2d6	Danio rerio
181330711	NP_001116708.1	hypothetical protein LOC554962	Danio rerio	100.0	100.0	99.6	0.0	NP_958499.1	DnaJ (Hsp40) homolog- subfamily A- member 3B	Danio rerio
181342116	NP_001116790.1	hypothetical protein LOC799807	Danio rerio	100.0	100.0	100.0	0.0	NP_001038328.1	crystallin- gamma M2d5	Danio rerio
181342116	NP_001116790.1	hypothetical protein LOC799807	Danio rerio	100.0	100.0	100.0	0.0	NP_001076408.1	crystallin- gamma family-like	Danio rerio

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
181342116	NP_001116790.1	hypothetical protein LOC799807	Danio rerio	100.0	100.0	99.4	0.0	NP_001103330.1	crystallin- gamma M2d10	Danio rerio
181342116	NP_001116790.1	hypothetical protein LOC799807	Danio rerio	100.0	100.0	98.8	0.0	NP_001073530.1	crystallin- gamma M2d4	Danio rerio
181342116	NP_001116790.1	hypothetical protein LOC799807	Danio rerio	100.0	100.0	98.8	0.0	NP_001038572.1	crystallin- gamma M2d7	Danio rerio
187607742	NP_001119951.1	hypothetical protein LOC100006144	Danio rerio	100.0	100.0	98.7	0.6	NP_001119955.1	novel protein similar to vertebrate pim oncogene family	Danio rerio
187608406	NP_001119919.1	hypothetical protein LOC791769	Danio rerio	100.0	100.0	99.4	0.0	NP_001005300.1	tRNA-specific adenosine deaminase-like protein 3	Danio rerio
189521959	XP_001923479.1	PREDICTED: hypothetical protein LOC100151290	Danio rerio	100.0	100.0	98.4	0.0	NP_001020667.1	novel immune-type receptor 12	Danio rerio
189523693	XP_001340168.2	PREDICTED: hypothetical protein LOC799853	Danio rerio	100.0	100.0	98.1	0.4	XP_001923016.1	novel protein (zgc:123060)	Danio rerio
190358598	NP_001121894.1	hypothetical protein LOC100151595	Danio rerio	100.0	100.0	99.7	0.0	NP_001002564.1	general transcription factor IIH- polypeptide 3	Danio rerio
192451469	NP_001122291.1	hypothetical protein LOC100149258	Danio rerio	100.0	100.0	99.4	0.0	NP_001014826.1	erbB-3b	Danio rerio
192453530	NP_001122292.1	hypothetical protein LOC100149563	Danio rerio	99.5	99.5	98.5	0.0	NP_001103640.1	serine protease	Xenopus (Silurana) tropicalis
192455632	NP_001122294.1	hypothetical protein LOC100149927	Danio rerio	100.0	100.0	99.4	0.0	NP_001076512.1	trace amine associated receptor 13e	Danio rerio
194578835	NP_001124143.1	hypothetical protein LOC100170837	Danio rerio	100.0	100.0	99.1	0.0	NP_001076376.1	trace amine associated receptor 12f	Danio rerio
198282015	NP_001103581.1	hypothetical protein LOC564145	Danio rerio	100.0	100.0	99.2	0.0	NP_001128340.2	Si:dkeyp-98a7.5 protein	Danio rerio
215422313	NP_001135847.1	hypothetical protein LOC797338	Danio rerio	100.0	100.0	99.8	0.0	NP_001038379.1	tripartite motif-containing 8	Danio rerio
224589090	NP_001139177.1	hypothetical protein LOC100003903	Danio rerio	100.0	100.0	100.0	0.0	NP_066361.1	ras-related protein Rap-2a precursor	Homo sapiens
224589090	NP_001139177.1	hypothetical protein LOC100003903	Danio rerio	100.0	100.0	100.0	0.0	NP_001155350.1	ras-related protein Rap-2a	Ovis aries

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
224589090	NP_001139177.1	hypothetical protein LOC100003903	Danio rerio	100.0	100.0	100.0	0.0	NP_001030288.1	RAP2A- member of RAS oncogene family	Xenopus (Silurana) tropicalis
224589090	NP_001139177.1	hypothetical protein LOC100003903	Danio rerio	100.0	100.0	100.0	0.0	NP_001080715.1	Rap2A GTPase	Xenopus laevis
237858566	NP_001153834.1	hypothetical protein LOC564145	Danio rerio	100.0	100.0	100.0	0.0	NP_001103581.1	novel rhamnose binding lectin-like	Danio rerio
237858570	NP_001128340.2	hypothetical protein LOC798039	Danio rerio	100.0	100.0	99.2	0.0	NP_001103581.1	novel rhamnose binding lectin-like	Danio rerio
237858617	NP_001153845.1	hypothetical protein LOC798111	Danio rerio	100.0	100.0	99.2	0.0	NP_001103581.1	novel rhamnose binding lectin-like	Danio rerio
292615801	XP_002662802.1	PREDICTED: hypothetical protein LOC100330879 isoform 2	Danio rerio	100.0	100.0	98.4	0.0	NP_938161.1	novel immune-type receptor 1l	Danio rerio
292617726	XP_002663436.1	PREDICTED: hypothetical protein LOC100330905	Danio rerio	100.0	100.0	99.8	0.0	NP_001107105.1	zinc finger-like gene 1	Danio rerio
292621532	XP_002664678.1	PREDICTED: hypothetical protein LOC100330956	Danio rerio	100.0	100.0	99.7	0.0	NP_571015.2	bonnie and clyde	Danio rerio
292624398	XP_002665637.1	PREDICTED: hypothetical protein LOC100330546	Danio rerio	100.0	100.0	98.7	0.2	NP_001093875.1	naked cuticle homolog 3	Danio rerio
312032382	NP_957141.2	hypothetical protein LOC393820	Danio rerio	100.0	100.0	100.0	0.0	NP_001187880.1	myeloma overexpressed gene 2 protein-like protein	Ictalurus punctatus
324710988	NP_001191325.1	hypothetical protein LOC100533191	Danio rerio	100.0	100.0	98.7	0.0	NP_001129461.1	zinc finger protein	Danio rerio
326669242	XP_002662820.2	PREDICTED: hypothetical protein LOC100332751 isoform 2	Danio rerio	100.0	100.0	99.7	0.0	NP_571731.1	novel immune-type receptor 4a isoform 3	Danio rerio
326673594	XP_003199934.1	PREDICTED: hypothetical protein LOC100538026	Danio rerio	100.0	100.0	98.4	0.0	NP_001164211.1	interleukin 4	Danio rerio
326679920	XP_003201411.1	PREDICTED: hypothetical protein LOC100537990	Danio rerio	100.0	100.0	99.4	0.0	NP_571912.1	CCAAT/enhancer binding protein (C/EBP) 1	Danio rerio
17511133	NP_491357.1	hypothetical protein ZK973.3	Caenorhabditis elegans	100.0	100.0	98.5	0.2	XP_003104811.1	CRE-PDP-1 protein	

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
17534493	NP_494960.1	hypothetical protein F58A6.9	Caenorhabditis elegans	100.0	100.0	98.7	0.0	XP_003115535.1	CRE-MSP-152 protein	
17538079	NP_495154.1	hypothetical protein F58A6.9	Caenorhabditis elegans	100.0	100.0	98.7	0.0	XP_003115535.1	CRE-MSP-152 protein	
17552332	NP_498052.1	hypothetical protein C27F2.5	Caenorhabditis elegans	100.0	100.0	99.6	0.0	XP_003106046.1	CRE-VPS-22 protein	
71983429	NP_001021221.1	hypothetical protein C46F11.2	Caenorhabditis elegans	100.0	100.0	98.7	0.0	XP_003113226.1	CRE-GSR-1 protein	
71988325	NP_001023196.1	hypothetical protein F38E11.6	Caenorhabditis elegans	100.0	100.0	99.5	0.5	NP_001023197.1	C. elegans protein F38E11.6b- confirmed by transcript evidence	
71990464	NP_001022749.1	hypothetical protein T05G5.9	Caenorhabditis elegans	100.0	100.0	99.7	0.3	NP_001022748.1	C. elegans protein T05G5.9a- confirmed by transcript evidence	
71993445	NP_496329.2	hypothetical protein R06F6.8	Caenorhabditis elegans	100.0	100.0	99.8	0.1	NP_496328.2	C. elegans protein R06F6.8b- partially confirmed by transcript evidence	
71997211	NP_498858.2	hypothetical protein ZK353.1	Caenorhabditis elegans	100.0	100.0	99.7	0.0	XP_003104122.1	CRE-CYY-1 protein	
71997217	NP_498857.2	hypothetical protein ZK353.1	Caenorhabditis elegans	100.0	100.0	99.2	0.6	XP_003104122.1	CRE-CYY-1 protein	
71999801	NP_001023622.1	hypothetical protein ZK795.4	Caenorhabditis elegans	100.0	100.0	100.0	0.0	NP_498790.1	SyNaptoBrevin related family member (snb-5)	
72000167	NP_506172.2	hypothetical protein R11D1.1	Caenorhabditis elegans	100.0	100.0	99.8	0.2	NP_506171.2	C. elegans protein R11D1.1b- confirmed by transcript evidence	
72001196	NP_503693.2	hypothetical protein Y45G12C.10	Caenorhabditis elegans	99.4	100.0	100.0	0.0	NP_503666.1	Seven TM Receptor family member (str-119)	
115532990	NP_001041015.1	hypothetical protein Y116A8C.28	Caenorhabditis elegans	100.0	100.0	98.1	0.0	XP_003103494.1	CRE-BCA-2 protein	
17505777	NP_491990.1	hypothetical protein C30F12.6	Caenorhabditis elegans	100.0	100.0	98.5	0.5	XP_003099070.1	CRE-NMUR-4 protein	
17507315	NP_492334.1	hypothetical protein F43G9.5	Caenorhabditis elegans	100.0	100.0	98.7	0.0	XP_003112093.1	CRE-CFIM-1 protein	
17510261	NP_492903.1	hypothetical protein Y53H1B.6	Caenorhabditis elegans	191.2	100.0	100.0	0.0	NP_872047.1	C. elegans protein Y51H1A.2c- confirmed by transcript evidence	
17539744	NP_502290.1	hypothetical protein F11A10.2	Caenorhabditis elegans	100.0	100.0	99.5	0.0	XP_003100775.1	CRE-REPO-1 protein	

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
17542418	NP_501739.1	hypothetical protein T13F2.12	Caenorhabditis elegans	99.1	99.1	100.0	0.0	NP_501782.1	Sperm-Specific family-class P family member (ssp-32)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	100.0	0.0	NP_494858.1	Major Sperm Protein family member (msp-3)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	99.2	0.0	NP_501781.1	Major Sperm Protein family member (msp-79)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	99.2	0.0	NP_494888.1	Major Sperm Protein family member (msp-33)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	99.2	0.0	NP_494901.1	Major Sperm Protein family member (msp-152)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	99.2	0.0	NP_494970.1	Major Sperm Protein family member (msp-49)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	99.2	0.0	NP_495143.1	Major Sperm Protein family member (msp-63)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	99.2	0.0	NP_500711.1	Major Sperm Protein family member (msp-57)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	98.4	0.0	NP_501849.1	Major Sperm Protein family member (msp-38)	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	100.0	0.0	NP_494898.1	CRE-MSP-142 protein	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	100.0	0.0	NP_501742.1	CRE-MSP-78 protein	
17543834	NP_501464.1	hypothetical protein Y59H11AM.1	Caenorhabditis elegans	100.0	100.0	98.4	0.0	XP_003092413.1	CRE-MSP-74 protein	
17555180	NP_499258.1	hypothetical protein T20G5.8	Caenorhabditis elegans	100.0	100.0	98.3	0.0	XP_003113097.1	CRE-DOD-6 protein	
17558698	NP_504130.1	hypothetical protein C49G7.3	Caenorhabditis elegans	100.0	100.0	98.8	0.0	NP_504129.1	PHAryngeal gland Toxin-related family member (phat-3)	
25150450	NP_741281.1	hypothetical protein T05D4.1	Caenorhabditis elegans	100.0	100.0	98.4	0.0	XP_003102129.1	CRE-ALDO-1 protein	

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
13384686	NP_079595.1	hypothetical protein LOC66050	Mus musculus	100.0	100.0	98.6	0.0	NP_001124790.1	trafficking protein particle complex subunit 2	Pongo abelii
13386026	NP_080804.1	hypothetical protein LOC68045	Mus musculus	100.0	100.0	98.4	0.4	NP_001030357.1	homeobox prox 1	Bos taurus
29244446	NP_808522.1	hypothetical protein LOC329641	Mus musculus	100.0	100.0	98.1	0.0	NP_001180660.1	chromosome 13 open reading frame 36	Macaca mulatta
30425270	NP_780720.1	hypothetical protein LOC241303	Mus musculus	100.0	100.0	98.9	0.0	NP_001181256.1	family with sequence similarity 78- member A	Macaca mulatta
30520227	NP_848879.1	family with sequence similarity 168- member A	Mus musculus	100.0	100.0	99.6	0.0	NP_001181715.1	family with sequence similarity 168- member A	Macaca mulatta
46559430	NP_941068.1	hypothetical protein LOC381714	Mus musculus	100.0	100.0	98.4	0.0	NP_892002.1	spermatogenesis associated glutamate (E)-rich protein 4-like	Mus musculus
47894404	NP_001001493.1	hypothetical protein LOC414077	Mus musculus	100.0	100.0	98.1	0.0	NP_001001645.1	CGI-140	Sus scrofa
49355814	NP_081357.2	hypothetical protein LOC69440	Mus musculus	100.0	100.0	99.1	0.0	NP_001102219.1	family with sequence similarity 116- member B	Rattus norvegicus
56090469	NP_001007580.1	hypothetical protein LOC329986	Mus musculus	100.0	100.0	98.1	0.0	NP_001034684.1	pramel family member	Mus musculus
58037527	NP_084373.1	hypothetical protein LOC78414	Mus musculus	100.0	100.0	99.1	0.0	NP_001032567.1	zinc finger protein 474 isoform 2	Bos taurus
62510085	NP_001007582.1	hypothetical protein LOC381406	Mus musculus	100.0	100.0	99.2	0.0	NP_076304.2	TP53-regulating kinase	Mus musculus
70608196	NP_001020431.1	histone cluster 2 family member	Mus musculus	100.0	100.0	99.0	0.0	NP_001079006.1	novel protein similar to histone H2a	Mus musculus
70794816	NP_001020559.1	hypothetical protein LOC433182	Mus musculus	100.0	100.0	98.4	0.0	NP_776474.2	alpha-enolase	Bos taurus
70794816	NP_001020559.1	hypothetical protein LOC433182	Mus musculus	100.0	100.0	98.8	0.0	NP_001419.1	alpha-enolase isoform 1	Homo sapiens
70794816	NP_001020559.1	hypothetical protein LOC433182	Mus musculus	100.0	100.0	98.8	0.0	NP_001182540.1	enolase 1- (alpha)	Macaca mulatta
70794816	NP_001020559.1	hypothetical protein LOC433182	Mus musculus	100.0	100.0	98.8	0.0	NP_001126461.1	alpha-enolase	Pongo abelii
70794816	NP_001020559.1	hypothetical protein LOC433182	Mus musculus	100.0	100.0	99.3	0.0	NP_036686.2	alpha-enolase isoform 2	Rattus norvegicus
70909314	NP_001020745.1	hypothetical protein LOC433492	Mus musculus	100.0	100.0	99.2	0.0	NP_780376.2	vomeromodulin precursor	Mus musculus
82880852	XP_924203.1	PREDICTED: hypothetical protein LOC71386	Mus musculus	100.0	100.0	100.0	0.0	XP_901600.1	mCG10536	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
82890520	XP_909759.1	PREDICTED: hypothetical protein LOC630994	Mus musculus	100.0	100.0	100.0	0.0	NP_001034683.1	late cornified envelope 3A	Mus musculus
82895170	XP_902736.1	PREDICTED: hypothetical protein LOC74989	Mus musculus	100.0	100.0	100.0	0.0	XP_927826.1	mCG54490	Mus musculus
82955329	XP_919864.1	PREDICTED: hypothetical protein LOC75272	Mus musculus	100.0	100.0	100.0	0.0	XP_127749.1	mCG10597	Mus musculus
83776567	NP_001033008.1	hypothetical protein LOC623898	Mus musculus	100.0	100.0	99.4	0.0	NP_001170981.1	spermatogenesis associated glutamate (E)-rich protein-like protein	Mus musculus
94377187	XP_992967.1	PREDICTED: hypothetical protein LOC76224	Mus musculus	100.0	100.0	100.0	0.0	XP_999732.1	mCG147461	Mus musculus
94378076	XP_001002242.1	PREDICTED: hypothetical protein LOC667035	Mus musculus	100.0	100.0	98.1	0.0	NP_956796.1	ubiquitin-40S ribosomal protein S27a	Ictalurus punctatus
94406152	XP_994008.1	PREDICTED: hypothetical protein LOC75521	Mus musculus	100.0	100.0	100.0	0.0	XP_997154.1	mCG148475	Mus musculus
94407493	XP_990452.1	PREDICTED: hypothetical protein LOC70936	Mus musculus	100.0	100.0	100.0	0.0	XP_998847.1	mCG14882	Mus musculus
110625692	NP_001001334.2	hypothetical protein LOC381350	Mus musculus	100.0	100.0	99.4	0.0	NP_001099595.1	sperm associated antigen 6-like	Rattus norvegicus
126032331	NP_001075117.1	hypothetical protein LOC434864	Mus musculus	100.0	100.0	98.7	0.0	NP_001107855.1	leucine zipper protein 4	Mus musculus
139948818	NP_598508.2	hypothetical protein LOC72722	Mus musculus	100.0	100.0	98.3	0.0	NP_001014095.1	family with sequence similarity 98- member A	Rattus norvegicus
147898821	NP_001079022.1	hypothetical protein LOC668963	Mus musculus	100.0	100.0	99.2	0.0	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus
147901510	NP_001079002.1	hypothetical protein LOC545652	Mus musculus	100.0	100.0	100.0	0.0	NP_001155080.1	interferon zeta	Mus musculus
147901510	NP_001079002.1	hypothetical protein LOC545652	Mus musculus	100.0	100.0	100.0	0.0	NP_922871.1	interferon zeta	Mus musculus
147901510	NP_001079002.1	hypothetical protein LOC545652	Mus musculus	100.0	100.0	99.5	0.0	NP_001079001.1	interferon zeta-like	Mus musculus
147905915	NP_001079010.1	hypothetical protein LOC626995	Mus musculus	100.0	100.0	99.2	0.0	NP_001079009.1	XPRAME family member 17	Mus musculus
148222126	NP_001079023.1	hypothetical protein LOC668964	Mus musculus	100.0	100.0	99.4	0.0	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
148222126	NP_001079023.1	hypothetical protein LOC668964	Mus musculus	100.0	100.0	99.4	0.0	NP_001079022.1	novel protein similar to germ cell-less homolog 1 (Drosophila) Gmcl1	Mus musculus
148222820	NP_001079014.1	hypothetical protein LOC630022	Mus musculus	100.0	100.0	99.6	0.0	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus
148224566	NP_001078994.1	hypothetical protein LOC434727	Mus musculus	100.0	100.0	100.0	0.0	NP_001078993.1	mCG1044706	Mus musculus
148228104	NP_001079020.1	hypothetical protein LOC668958	Mus musculus	100.0	100.0	99.6	0.0	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus
148233024	NP_001079012.1	hypothetical protein LOC627264	Mus musculus	100.0	100.0	99.4	0.0	NP_001079014.1	novel protein similar to human germ cell-less homolog 1 (Drosophila)-like (GMCL1L)	Mus musculus
148233024	NP_001079012.1	hypothetical protein LOC627264	Mus musculus	100.0	100.0	99.2	0.0	NP_001079021.1	novel protein similar to human germ cell-less homolog 1 (Drosophila)-like (GMCL1L)	Mus musculus
148233024	NP_001079012.1	hypothetical protein LOC627264	Mus musculus	100.0	100.0	99.0	0.0	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus
148235086	NP_001078992.1	hypothetical protein LOC434725	Mus musculus	100.0	100.0	99.4	0.0	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus
148235545	NP_001091446.1	hypothetical protein LOC433486	Mus musculus	100.0	100.0	99.2	0.0	NP_941057.1	sperm motility kinase X	Mus musculus
148235915	NP_001079021.1	hypothetical protein LOC668960	Mus musculus	100.0	100.0	99.4	0.0	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus
148276987	NP_001087229.1	hypothetical protein LOC66975 isoform 3	Mus musculus	100.0	100.0	98.8	0.2	NP_001128904.1	DKFZP459P083 protein	Pongo abelii
149234190	XP_001474281.1	PREDICTED: hypothetical protein LOC100040213 isoform 1	Mus musculus	100.0	100.0	99.3	0.0	NP_291094.2	component of Sp100-rs	Mus musculus
149234190	XP_001474281.1	PREDICTED: hypothetical protein LOC100040213 isoform 1	Mus musculus	100.0	100.0	98.3	0.0	NP_001075215.1	component of Sp100-rs-like	Mus musculus
149247823	XP_001477761.1	PREDICTED: hypothetical protein LOC100041903 isoform 1	Mus musculus	100.0	100.0	99.3	0.0	NP_291094.2	component of Sp100-rs	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
149247823	XP_001477761.1	PREDICTED: hypothetical protein LOC100041903 isoform 1	Mus musculus	100.0	100.0	98.3	0.0	NP_001075215.1	component of Sp100-rs-like	Mus musculus
149250453	XP_001479087.1	hypothetical protein LOC75485	Mus musculus	100.0	100.0	100.0	0.0	NP_083584.1	novel protein	Mus musculus
149252547	XP_001472259.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	99.1	0.0	NP_001128148.1	major urinary protein 8	Mus musculus
149252547	XP_001472259.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	99.6	0.0	NP_001156483.1	major urinary protein 1 isoform a	Mus musculus
149252547	XP_001472259.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	99.1	0.0	NP_001128147.1	major urinary protein 7	Mus musculus
149252547	XP_001472259.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	99.1	0.0	NP_001128146.1	major urinary protein 13	Mus musculus
149253374	XP_001472163.1	PREDICTED: hypothetical protein LOC78723	Mus musculus	100.0	100.0	100.0	0.0	XP_001476787.1	mCG148017	Mus musculus
149254210	XP_001473822.1	PREDICTED: hypothetical protein LOC75610	Mus musculus	100.0	100.0	100.0	0.0	XP_001475763.1	mCG1041431	Mus musculus
149255846	XP_001480716.1	PREDICTED: hypothetical protein LOC74269	Mus musculus	100.0	100.0	100.0	0.0	XP_001481062.1	mCG1036725	Mus musculus
149257976	XP_001477453.1	PREDICTED: hypothetical protein LOC76679	Mus musculus	100.0	100.0	100.0	0.0	XP_001478810.1	mCG147567	Mus musculus
149258434	XP_001473395.1	PREDICTED: hypothetical protein LOC665044	Mus musculus	100.0	100.0	100.0	0.0	XP_979245.1	mCG147749	Mus musculus
149260981	XP_001477810.1	PREDICTED: hypothetical protein LOC634517	Mus musculus	100.0	100.0	98.7	0.0	NP_001095100.1	keratin associated protein 10-like	Mus musculus
149260983	XP_001477827.1	PREDICTED: hypothetical protein LOC100041281	Mus musculus	100.0	100.0	98.3	0.0	NP_001095100.1	keratin associated protein 10-like	Mus musculus
149264211	XP_001473814.1	PREDICTED: hypothetical protein LOC442803	Mus musculus	100.0	100.0	100.0	0.0	XP_001474878.1	RIKEN cDNA A830005F24	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
149264499	XP_001474052.1	PREDICTED: hypothetical protein LOC77166	Mus musculus	100.0	100.0	100.0	0.0	XP_001479007.1	mCG54345	Mus musculus
149265010	XP_001476831.1	PREDICTED: hypothetical protein LOC100041664	Mus musculus	99.5	100.0	98.5	0.0	NP_001019877.2	alpha takusan-like	Mus musculus
149265071	XP_001477556.1	PREDICTED: hypothetical protein LOC100041970 isoform 1	Mus musculus	100.0	100.0	98.0	0.0	NP_001019877.2	alpha takusan-like	Mus musculus
149267858	XP_001479073.1	PREDICTED: hypothetical protein LOC75531	Mus musculus	100.0	100.0	100.0	0.0	XP_001481001.1	mCG146828	Mus musculus
149271680	XP_001473919.1	PREDICTED: hypothetical protein LOC75509	Mus musculus	100.0	100.0	99.1	0.0	NP_001092312.1	mCG15151	Mus musculus
149272284	XP_001475702.1	PREDICTED: hypothetical protein LOC71248	Mus musculus	100.0	100.0	100.0	0.0	XP_001473842.1	novel aspartic/glutamic acid-rich region containing protein	Mus musculus
149272541	XP_001472171.1	hypothetical protein LOC100039890	Mus musculus	100.0	100.0	98.2	0.0	NP_001107855.1	leucine zipper protein 4	Mus musculus
149275217	XP_001473429.1	PREDICTED: hypothetical protein LOC100041579 isoform 2	Mus musculus	100.0	100.0	98.9	0.0	NP_001157516.1	C-C motif chemokine 27 isoform 3	Mus musculus
150010627	NP_001092796.1	hypothetical protein LOC100042782	Mus musculus	100.0	100.0	100.0	0.0	NP_112551.2	ferritin heavy polypeptide-like 17	Mus musculus
150010669	NP_001092791.1	pramel3 family member	Mus musculus	98.9	100.0	99.4	0.0	NP_113567.1	preferentially expressed antigen in melanoma-like 3	Mus musculus
153791512	NP_001093390.1	hypothetical protein LOC100039890	Mus musculus	100.0	100.0	98.2	0.0	NP_001107855.1	leucine zipper protein 4	Mus musculus
153791609	NP_001093113.1	hypothetical protein LOC100042254	Mus musculus	100.0	100.0	99.0	0.4	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus
153792541	NP_001092817.1	hypothetical protein LOC100042175	Mus musculus	100.0	100.0	98.1	0.0	NP_001129948.1	Sycp3 like X-linked	Mus musculus
154759297	NP_001094080.1	hypothetical protein LOC100042175	Mus musculus	100.0	100.0	98.1	0.0	NP_001129948.1	Sycp3 like X-linked	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
155969725	NP_001095126.1	hypothetical protein LOC100043123	Mus musculus	100.0	100.0	99.6	0.0	NP_954671.1	CMRF35-like molecule 3 precursor	Mus musculus
156119549	NP_001094893.1	hypothetical protein LOC100048775	Mus musculus	100.0	100.0	100.0	0.0	NP_041266.1	small t-antigen	Murine polyomavirus
156713479	NP_001096147.1	hypothetical protein LOC100038977	Mus musculus	100.0	100.0	98.1	0.0	NP_001129948.1	Sycp3 like X-linked	Mus musculus
156938235	NP_001028295.2	hypothetical protein LOC13999	Mus musculus	100.0	100.0	100.0	0.0	NP_001170871.1	ethanol induced 1 isoform 2	Mus musculus
158303280	NP_001094079.1	hypothetical protein LOC100042175	Mus musculus	100.0	100.0	98.1	0.0	NP_001129948.1	Sycp3 like X-linked	Mus musculus
158303288	NP_001096148.1	hypothetical protein LOC100042175	Mus musculus	100.0	100.0	98.1	0.0	NP_001129948.1	Sycp3 like X-linked	Mus musculus
158303298	NP_001093389.1	hypothetical protein LOC100042175	Mus musculus	100.0	100.0	98.1	0.0	NP_001129948.1	Sycp3 like X-linked	Mus musculus
159032007	NP_001103720.1	hypothetical protein LOC100042175	Mus musculus	100.0	100.0	98.1	0.0	NP_001129948.1	Sycp3 like X-linked	Mus musculus
160707965	NP_899143.2	pramel3 family member	Mus musculus	99.8	100.0	99.6	0.0	NP_113567.1	preferentially expressed antigen in melanoma-like 3	Mus musculus
160707969	NP_808586.1	pramel3 family member	Mus musculus	98.9	100.0	99.1	0.0	NP_113567.1	preferentially expressed antigen in melanoma-like 3	Mus musculus
168480088	NP_001094976.1	hypothetical protein LOC381582	Mus musculus	100.0	100.0	98.8	0.0	NP_001193892.1	transmembrane protein C1orf70	Bos taurus
168480088	NP_001094976.1	hypothetical protein LOC381582	Mus musculus	100.0	100.0	98.8	0.0	NP_001108220.1	transmembrane protein C1orf70	Homo sapiens
169808401	NP_001116132.1	hypothetical protein LOC100042314	Mus musculus	100.0	100.0	99.1	0.0	NP_032208.2	glutathione S-transferase A2	Mus musculus
169808420	NP_001116134.1	hypothetical protein LOC100039042	Mus musculus	100.0	100.0	99.3	0.0	NP_001013846.1	Eif1a-like	Mus musculus
170172505	NP_001116207.1	hypothetical protein LOC382244	Mus musculus	100.0	100.0	98.7	0.0	NP_001107855.1	leucine zipper protein 4	Mus musculus
226442863	NP_001092803.3	hypothetical protein LOC100040606	Mus musculus	100.0	100.0	99.4	0.0	NP_082231.2	germ cell-less protein-like 1-like	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
236461352	NP_001153601.1	hypothetical protein LOC100040786	Mus musculus	100.0	100.0	98.2	0.0	NP_001017394.2	spermiogenesis specific transcript on the Y family member	Mus musculus
236461352	NP_001153601.1	hypothetical protein LOC100040786	Mus musculus	100.0	100.0	98.2	0.0	NP_076035.3	spermiogenesis specific transcript on the Y 2	Mus musculus
236461435	NP_001153603.1	hypothetical protein LOC100040911	Mus musculus	100.0	100.0	98.7	0.0	NP_001017394.2	spermiogenesis specific transcript on the Y family member	Mus musculus
236462435	NP_001153607.1	hypothetical protein LOC100039574	Mus musculus	100.0	100.0	98.7	0.0	NP_001017394.2	spermiogenesis specific transcript on the Y family member	Mus musculus
236462435	NP_001153607.1	hypothetical protein LOC100039574	Mus musculus	100.0	100.0	98.7	0.0	NP_076035.3	spermiogenesis specific transcript on the Y 2	Mus musculus
236462500	NP_001153608.1	hypothetical protein LOC100042428	Mus musculus	100.0	100.0	98.7	0.0	NP_001017394.2	spermiogenesis specific transcript on the Y family member	Mus musculus
236462859	NP_001153609.1	hypothetical protein LOC100039614	Mus musculus	100.0	100.0	98.7	0.0	NP_001153613.1	putative	Mus musculus
236462859	NP_001153609.1	hypothetical protein LOC100039614	Mus musculus	100.0	100.0	98.7	0.0	NP_001153603.1	putative	Mus musculus
236462859	NP_001153609.1	hypothetical protein LOC100039614	Mus musculus	100.0	100.0	98.2	0.0	NP_001017394.2	spermiogenesis specific transcript on the Y family member	Mus musculus
240849195	NP_001155836.1	hypothetical protein LOC100039120	Mus musculus	100.0	100.0	98.1	0.0	NP_001129948.1	Sycp3 like X-linked	Mus musculus
254540154	NP_081195.1	hypothetical protein LOC69038 isoform 1	Mus musculus	100.0	100.0	100.0	0.0	NP_001181992.1	protein NEF1	Bos taurus
254540154	NP_081195.1	hypothetical protein LOC69038 isoform 1	Mus musculus	100.0	100.0	100.0	0.0	NP_001187759.1	NEF1 protein	Ictalurus punctatus
254540154	NP_081195.1	hypothetical protein LOC69038 isoform 1	Mus musculus	100.0	100.0	100.0	0.0	NP_001005138.1	chromosome 11 open reading frame 10	Xenopus laevis
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	99.1	0.0	NP_001070036.1	ribose-phosphate pyrophosphokinase 1	Danio rerio

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	98.7	0.0	NP_998698.1	phosphoribosyl pyrophosphate synthetase 1A isoform 2	Danio rerio
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	98.4	0.0	NP_080938.1	ribose-phosphate pyrophosphokinase 2	Mus musculus
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	98.4	0.0	NP_002756.1	ribose-phosphate pyrophosphokinase 2	Pongo abelii
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	99.7	0.0	NP_001009694.1	ribose-phosphate pyrophosphokinase I -like	Rattus norvegicus
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	98.4	0.0	NP_036766.1	ribose-phosphate pyrophosphokinase 2	Rattus norvegicus
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	98.1	0.0	NP_001099148.1	phosphoribosyl pyrophosphate synthetase 1-like 1	Rattus norvegicus
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	98.4	0.6	NP_001133837.1	Ribose-phosphate pyrophosphokinase 1	Salmo salar
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	98.7	0.0	NP_989140.1	phosphoribosyl pyrophosphate synthetase 1	Xenopus (Silurana) tropicalis
256418956	NP_001032835.2	hypothetical protein LOC328099	Mus musculus	100.0	100.0	98.7	0.0	NP_001083991.1	phosphoribosyl pyrophosphate synthetase 1	Xenopus laevis
257153362	NP_001158052.1	hypothetical protein LOC233812	Mus musculus	100.0	100.0	98.2	0.0	NP_001134691.1	CP52B protein	Salmo salar
257153362	NP_001158052.1	hypothetical protein LOC233812	Mus musculus	100.0	100.0	99.4	0.0	NP_001085337.1	chromosome 16 open reading frame 52	Xenopus laevis
262263351	NP_001160118.1	hypothetical protein LOC100040867	Mus musculus	100.0	100.0	100.0	0.0	NP_083457.1	Slx-like 1	Mus musculus
262263351	NP_001160118.1	hypothetical protein LOC100040867	Mus musculus	100.0	100.0	99.4	0.0	NP_001207426.1	Slx-like	Mus musculus
268837834	NP_001013842.2	hypothetical protein LOC434674	Mus musculus	100.0	100.0	99.1	0.0	NP_795976.1	integral membrane transport protein UST1R	Mus musculus
309262203	XP_003085746.1	PREDICTED: hypothetical protein LOC100041281	Mus musculus	100.0	100.0	98.7	0.0	NP_001095100.1	keratin associated protein 10-like	Mus musculus
309263531	XP_003086052.1	PREDICTED: hypothetical protein LOC100040214	Mus musculus	100.0	100.0	98.7	0.0	NP_570926.1	keratin-associated protein 16-8	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
309263683	XP_003086108.1	PREDICTED: hypothetical protein LOC100502931	Mus musculus	100.0	100.0	100.0	0.0	XP_003085223.1	novel protein	Mus musculus
309263722	XP_003086121.1	PREDICTED: hypothetical protein LOC76651	Mus musculus	100.0	100.0	100.0	0.0	XP_001480169.1	mCG53587	Mus musculus
309264806	XP_924015.3	PREDICTED: hypothetical protein LOC630971	Mus musculus	100.0	100.0	99.0	0.0	NP_001018089.1	late cornified envelope protein	Mus musculus
309264808	XP_003086361.1	PREDICTED: hypothetical protein LOC69514	Mus musculus	100.0	100.0	99.0	0.0	NP_001018089.1	late cornified envelope protein	Mus musculus
309266974	XP_915666.3	PREDICTED: hypothetical protein LOC546397	Mus musculus	100.0	100.0	98.3	0.0	XP_001474005.2	mCG114344	Mus musculus
309267466	XP_003084502.1	PREDICTED: hypothetical protein LOC100503095	Mus musculus	100.0	100.0	100.0	0.0	XP_003086258.1	unknown	Mus musculus
309267468	XP_003084503.1	PREDICTED: hypothetical protein LOC100503433	Mus musculus	100.0	100.0	100.0	0.0	NP_291094.2	component of Sp100-rs	Mus musculus
309267468	XP_003084503.1	PREDICTED: hypothetical protein LOC100503433	Mus musculus	100.0	100.0	99.0	0.0	NP_001075215.1	component of Sp100-rs-like	Mus musculus
309267470	XP_984257.3	PREDICTED: hypothetical protein LOC665317 isoform 1	Mus musculus	100.0	100.0	99.0	0.0	NP_291094.2	component of Sp100-rs	Mus musculus
309267543	XP_003084517.1	PREDICTED: hypothetical protein LOC329307	Mus musculus	100.0	100.0	100.0	0.0	XP_003086240.1	mCG147437	Mus musculus
309267576	XP_003084532.1	PREDICTED: hypothetical protein LOC791270	Mus musculus	100.0	100.0	100.0	0.0	XP_003086280.1	mCG147252	Mus musculus
309267637	XP_003084543.1	PREDICTED: hypothetical protein LOC70730	Mus musculus	100.0	100.0	100.0	0.0	XP_003086277.1	novel protein	Mus musculus
309267855	XP_003084568.1	PREDICTED: hypothetical protein LOC100504707	Mus musculus	100.0	100.0	100.0	0.0	XP_003086329.1	mCG17500- isoform CRA_c	Mus musculus
309267979	XP_001478629.2	PREDICTED: hypothetical protein LOC630994	Mus musculus	100.0	100.0	100.0	0.0	NP_001034683.1	late cornified envelope 3A	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
309267981	XP_001478635.2	PREDICTED: hypothetical protein LOC69514	Mus musculus	100.0	100.0	99.0	0.0	NP_001018089.1	late cornified envelope protein	Mus musculus
309268049	XP_003084590.1	PREDICTED: hypothetical protein LOC791377	Mus musculus	100.0	100.0	100.0	0.0	XP_003086381.1	mCG147415	Mus musculus
309268105	XP_003084619.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	99.4	0.6	NP_001186928.1	major urinary protein 14	Mus musculus
309268105	XP_003084619.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	98.9	0.6	NP_001128599.1	major urinary proteins 11 and 8	Mus musculus
309268105	XP_003084619.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	98.3	0.6	NP_001116119.1	major urinary protein 10	Mus musculus
309268105	XP_003084619.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	98.9	0.6	NP_032673.3	major urinary protein 2 precursor	Mus musculus
309268105	XP_003084619.1	PREDICTED: hypothetical protein LOC100039116	Mus musculus	100.0	100.0	98.9	0.6	NP_112465.2	major urinary protein 1 isoform b	Mus musculus
309268814	XP_978556.2	PREDICTED: hypothetical protein LOC27493 isoform 3	Mus musculus	100.0	100.0	100.0	0.0	XP_924274.2	polyQ-containing protein CAG-8	Mus musculus
309270090	XP_003085039.1	PREDICTED: hypothetical protein LOC100040697 isoform 2	Mus musculus	100.0	100.0	98.5	0.0	NP_001025101.2	alpha7-takusan-like	Mus musculus
309270090	XP_003085039.1	PREDICTED: hypothetical protein LOC100040697 isoform 2	Mus musculus	99.5	99.5	98.5	0.0	NP_001158199.1	alpha7-takusan	Mus musculus
309270097	XP_001475650.2	PREDICTED: hypothetical protein LOC100040797	Mus musculus	100.0	100.0	100.0	0.0	NP_796109.1	alpha28-takusan	Mus musculus
309270100	XP_001475897.2	PREDICTED: hypothetical protein LOC100040797	Mus musculus	100.0	100.0	100.0	0.0	NP_796109.1	alpha28-takusan	Mus musculus
309270162	XP_003085002.1	PREDICTED: hypothetical protein LOC666329	Mus musculus	100.0	100.0	98.0	0.0	NP_001158199.1	alpha7-takusan	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
309270162	XP_003085002.1	PREDICTED: hypothetical protein LOC666329	Mus musculus	99.5	99.5	98.0	0.0	NP_001025101.2	alpha7-takusan-like	Mus musculus
309270164	XP_990872.3	PREDICTED: hypothetical protein LOC666750	Mus musculus	100.0	100.0	98.6	0.0	NP_796109.1	alpha28-takusan	Mus musculus
309270171	XP_003085004.1	PREDICTED: hypothetical protein LOC100041515	Mus musculus	100.0	100.0	99.4	0.0	NP_001171186.1	alpha6-takusan-like isoform 2	Mus musculus
309270183	XP_003085007.1	PREDICTED: hypothetical protein LOC100042054 isoform 1	Mus musculus	99.5	99.5	100.0	0.0	NP_001158199.1	alpha7-takusan	Mus musculus
309270183	XP_003085007.1	PREDICTED: hypothetical protein LOC100042054 isoform 1	Mus musculus	100.0	100.0	99.0	0.0	NP_001025101.2	alpha7-takusan-like	Mus musculus
309270185	XP_003085009.1	PREDICTED: hypothetical protein LOC100042054 isoform 3	Mus musculus	99.5	99.5	98.1	1.9	NP_001158199.1	alpha7-takusan	Mus musculus
309270196	XP_003085012.1	PREDICTED: hypothetical protein LOC666684	Mus musculus	100.0	100.0	98.8	0.0	NP_001171186.1	alpha6-takusan-like isoform 2	Mus musculus
309270202	XP_003085014.1	PREDICTED: hypothetical protein LOC100503971 isoform 1	Mus musculus	100.0	100.0	99.5	0.0	NP_001158199.1	alpha7-takusan	Mus musculus
309270202	XP_003085014.1	PREDICTED: hypothetical protein LOC100503971 isoform 1	Mus musculus	99.5	99.5	98.5	0.0	NP_001025101.2	alpha7-takusan-like	Mus musculus
309270211	XP_001477388.2	PREDICTED: hypothetical protein LOC666750	Mus musculus	100.0	100.0	98.6	0.0	NP_796109.1	alpha28-takusan	Mus musculus
309270215	XP_003085019.1	PREDICTED: hypothetical protein LOC100041678 isoform 2	Mus musculus	99.5	99.5	99.0	0.0	NP_001158199.1	alpha7-takusan	Mus musculus
309270215	XP_003085019.1	PREDICTED: hypothetical protein LOC100041678 isoform 2	Mus musculus	100.0	100.0	98.0	0.0	NP_001025101.2	alpha7-takusan-like	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
309270224	XP_003085023.1	PREDICTED: hypothetical protein LOC100042054 isoform 3	Mus musculus	99.5	99.5	98.1	1.9	NP_001158199.1	alpha7-takusan	Mus musculus
309270233	XP_001477411.2	PREDICTED: hypothetical protein LOC666750	Mus musculus	100.0	100.0	98.6	0.0	NP_796109.1	alpha28-takusan	Mus musculus
309270237	XP_003085025.1	PREDICTED: hypothetical protein LOC100041530	Mus musculus	100.0	100.0	99.3	0.0	NP_001171185.1	alpha6-takusan-like isoform 1	Mus musculus
309270239	XP_001477450.2	PREDICTED: hypothetical protein LOC100040797	Mus musculus	100.0	100.0	100.0	0.0	NP_796109.1	alpha28-takusan	Mus musculus
309270243	XP_001477591.2	PREDICTED: hypothetical protein LOC666750	Mus musculus	100.0	100.0	98.6	0.0	NP_796109.1	alpha28-takusan	Mus musculus
309270262	XP_003085029.1	PREDICTED: hypothetical protein LOC100042100 isoform 2	Mus musculus	100.0	100.0	99.3	0.0	NP_001171185.1	alpha6-takusan-like isoform 1	Mus musculus
309270286	XP_001472902.2	PREDICTED: hypothetical protein LOC100039452	Mus musculus	100.0	100.0	98.6	0.0	NP_796109.1	alpha28-takusan	Mus musculus
309270304	XP_003085055.1	PREDICTED: hypothetical protein LOC100503880	Mus musculus	100.0	100.0	100.0	0.0	XP_003085954.1	RIKEN cDNA D830044D21	Mus musculus
309270808	XP_001474347.2	PREDICTED: hypothetical protein LOC100040214	Mus musculus	100.0	100.0	98.7	0.0	NP_570926.1	keratin-associated protein 16-8	Mus musculus
309270969	XP_003085173.1	PREDICTED: hypothetical protein LOC100502931	Mus musculus	100.0	100.0	100.0	0.0	XP_003085223.1	novel protein	Mus musculus
309271048	XP_003085187.1	PREDICTED: hypothetical protein LOC75039 isoform 2	Mus musculus	100.0	100.0	100.0	0.0	XP_003086125.1	mCG140970- isoform CRA_b	Mus musculus
309271050	XP_001481212.2	PREDICTED: hypothetical protein LOC75039 isoform 1	Mus musculus	100.0	100.0	100.0	0.0	XP_003086124.1	mCG140970- isoform CRA_a	Mus musculus
309271556	XP_001473158.2	PREDICTED: hypothetical protein LOC434863	Mus musculus	100.0	100.0	98.7	0.0	NP_001107855.1	leucine zipper protein 4	Mus musculus

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
309271612	XP_984464.3	PREDICTED: hypothetical protein LOC665338 isoform 2	Mus musculus	100.0	100.0	99.3	0.0	NP_291094.2	component of Sp100-rs	Mus musculus
309271612	XP_984464.3	PREDICTED: hypothetical protein LOC665338 isoform 2	Mus musculus	100.0	100.0	98.3	0.0	NP_001075215.1	component of Sp100-rs-like	Mus musculus
309272207	XP_003085496.1	PREDICTED: hypothetical protein LOC100503001	Mus musculus	100.0	100.0	100.0	0.0	XP_003086530.1	RIKEN cDNA 1700010H22	Mus musculus
309272680	XP_003085582.1	PREDICTED: hypothetical protein LOC100503923	Mus musculus	100.0	100.0	99.3	0.0	NP_291094.2	component of Sp100-rs	Mus musculus
309272680	XP_003085582.1	PREDICTED: hypothetical protein LOC100503923	Mus musculus	100.0	100.0	98.3	0.0	NP_001075215.1	component of Sp100-rs-like	Mus musculus
309272924	XP_003085636.1	PREDICTED: hypothetical protein LOC100504459	Mus musculus	100.0	100.0	100.0	0.0	XP_003085950.1	immune system released activating agent	Mus musculus
309272972	XP_003085643.1	PREDICTED: hypothetical protein LOC100504276 isoform 2	Mus musculus	100.0	100.0	98.4	0.0	NP_001107855.1	leucine zipper protein 4	Mus musculus
309272974	XP_003085642.1	hypothetical protein LOC100039890	Mus musculus	100.0	100.0	98.2	0.0	NP_001107855.1	leucine zipper protein 4	Mus musculus
312176356	NP_001185917.1	hypothetical protein LOC666184	Mus musculus	100.0	100.0	98.4	0.0	NP_001107855.1	leucine zipper protein 4	Mus musculus
313151172	NP_001186237.1	hypothetical protein LOC13999	Mus musculus	100.0	100.0	100.0	0.0	NP_001170871.1	ethanol induced 1 isoform 2	Mus musculus
148277000	NP_079217.2	hypothetical protein LOC80006 isoform 2	Homo sapiens	100.0	100.0	98.1	1.4	NP_001128904.1	DKFZP459P083 protein	Pongo abelii
148277004	NP_001087225.1	hypothetical protein LOC80006 isoform 3	Homo sapiens	100.0	100.0	99.5	0.0	NP_001128904.1	DKFZP459P083 protein	Pongo abelii
149944593	NP_056112.1	hypothetical protein LOC23349	Homo sapiens	100.0	100.0	98.8	0.0	NP_001128900.1	DKFZP459L2316 protein	Pongo abelii
169217554	XP_001722881.1	PREDICTED: hypothetical protein LOC100127885	Homo sapiens	100.0	100.0	100.0	0.0	XP_001721823.1	liver-related low express protein 1	Homo sapiens

Query GI	Query Ref	Query Name	Query Type	Percent Query Cover	Percent Subject Cover	Percent Similarity	Percent Gaps	Subject Ref	Subject Name	Subject Type
169218060	XP_001725001.1	PREDICTED: hypothetical protein LOC100127885	Homo sapiens	100.0	100.0	100.0	0.0	XP_001721823.1	liver-related low express protein 1	Homo sapiens
178057345	NP_001116640.1	hypothetical protein LOC729533	Homo sapiens	100.0	100.0	99.3	0.0	NP_001094380.1	amyloid-beta peptide-induced protein p17	Homo sapiens
254540132	NP_001156896.1	hypothetical protein LOC150678 isoform 2	Homo sapiens	100.0	100.0	100.0	0.0	NP_001187880.1	myeloma overexpressed gene 2 protein-like protein	Ictalurus punctatus
254540132	NP_001156896.1	hypothetical protein LOC150678 isoform 2	Homo sapiens	100.0	100.0	100.0	0.0	NP_001005143.1	myeloma overexpressed gene 2 protein homolog	Xenopus (Silurana) tropicalis
310119066	XP_003118565.1	PREDICTED: hypothetical protein LOC100506055	Homo sapiens	100.0	100.0	100.0	0.0	NP_940932.2	matrix-remodeling-associated protein 7 isoform 3	Homo sapiens
310132225	XP_003120900.1	PREDICTED: hypothetical protein LOC100287593	Homo sapiens	100.0	100.0	100.0	0.0	XP_002344133.1	B-cell growth factor	Homo sapiens
310132381	XP_003120926.1	PREDICTED: hypothetical protein LOC100510248	Homo sapiens	100.0	100.0	99.4	0.0	NP_114169.2	keratin-associated protein 9-8	Homo sapiens