Approaches for Gene and Genome Analysis

What is encoded by the sequence information?

What is the probable function in the cell?

(by Bettina Siebers)

By sequencing of a clone of the „whole genome shotgun library“ from the thermophilic Archaeon *Sulfolobus solfataricus the* following nucleotide sequence was identified. The emerging questions are if this region encodes for a certain gene (open reading frame, ORF) and if yes what the possible function of the gene product (i.e. protein) in *S. solfataricus* might be.

Nucleotide-sequence: (Sso_0987)

GAGTCCAGCTGAAGTACATTCCTATTTTAACGCCTTTGGATCCCCTATAGATGCGTTTATAAATGAATTT
TATGCAATATCGACATTTATACATGAAGAAATAAGTAAATTAAATATTAAGAATGAGCCTTATATATTCA
TATTAGGAGATAAGAGAGCGTCGATAGCTTGGAATGAAAAAGAGTTCATGGAAATTGTAATGAGAGATGA
AAGGTTTAAAACTCATTTGAAAAACTTAAGGCTGTGGTTAGGAAATGAAAAGGATTGAATCATTGTGGTT
ACCTGAGGACATTAAAAAAGTGTGGATGATAACATTTGAATTGCAAAAGATAGCAAGTGTTGGAGGATTA
GGAAATGCTGTATATAACATAGCTAAACATCTAGCTGAAAAGGGAGTAGATATTACGGTTTTCTTGCCAT
CTCATGGAAGGCATCTAAATGAATATTATAGGTCTCTCTTAAGTTTAAGACATATTGACATGATCGTTGA
GGGAAGAAGAAAGGGTATAGATAATAACTATTATAATTATAAAATAGGATTTGAAGAAGGAAAAATAGAT
AATTTCAAGGTAATTTTGGTAAAAGGATTAGATTATAACACTGGTAGGGTATTAGATTCATGGAATATCT
ATGATAACACAATGGAGAAAATTTCGCTATTAACAAGAGGATTAGAGGGGTTCACTTTAGGCAATCTATC
TAACCTTCCAGATATAATCCATGCACAAGATTGGCACGCCGTAATTCCCGCAGTTAGGATAAAACAACTC
TTGGAAGAGAGACGAATAATCGTCCCATTTATTTACACTATTCACTTACTGAATTACATTGGTGTACCTT
GGCACTACGCTTCTCAAGACTGGTCTGGAATTGAGGATTGTTGGCATTATATTTGGATGGTAGCTAAGCA
TGAATTATATAAATATTCATATGTATGGGATGTGTTGTCAGGTGGAAAAATTGAGAAATTTGGTTGTTAC
GAGGCTGATATGGTGAGTAGTGTAAGTTACAGTTATTTAAGTTTTGACGTATTTAATTTTGTAGGAAATT
GGGTAGCCAATAAATCATGTGTAACGTATAATGGTACGGATTGGGATGTGGAAGAGATCCAAAATAAGGC
TGTTACAATGTATGGAACTAAGGACAGAAGGGAGTTAAGGAGAAGGCTACTTTCATCTCTGCACTCCCTA
AGGGTTATTCCAGAGGATTACACTACCGGCAATATGCTATGGAATAATAGAAATAGACTAGGATTAAGAG
ATGATTGGACTTACGACGATTTAGGGGAAGGACCCCTTGTTCTATTCACTGGGAGATTAGTATATCAAAA
AGGTGTAGATTTACTTTTAAGGGCTATGAAAACAGTTGTGAATGAGATAAATAATGCTAGGCTCTTAATT
TTTGGGCTACCGTCTGGGGATTACAACTTACTTTGGGATATTATAGAGAGGGCTTCAGAAATTAAGGATA
ATATTAGGTTAATAGTGGGTAGAATGGATTTAGATTTATATAAATTGTTTCACTACGTATCTTCAGTCTT
CGTCATTCCATCTAGATGGGAACCATTTGGCATAAATTCAATTGAGGCTATGGCTATGGGGTTGCCAGTA
ATTGCCTATAGCGTAGGGGGTTTAAGAGAAACAGTCGTTGACATTAGGGAGGATAAGAATAATGCCACAG
GTCTCTTAATTAAGCCTGAAAGTATAGACGAATTAGCTAGAGCCATAAGGATTGCATTATATTTGTCAGA
GGCTTCTGAGCTAAATAAAAGCGATTTGTTATATAAGGCGAGTGAGGTTAAAGTAGATGATACGAGATAC
TGGGATAAGGTACGTGAAAATGCAATAACTAGGGTTAAGAGTAGATTTAGATGGGATGCAGTGATAAACT
CCTTAACTGAGTGTTATAGAAAGACTCTCGATATGGCTAAATATAGGGCTTTAGCCTCTTTTTGAGGAGA
TATCATGAATATAAAGAACGCATTGGAAAGAAAGGACGTCAAATATCTGATAAGGAATATAAGAAGTTTA
CTCAACTTGCTAAAAAGTAAAGATGGATTAGAAAGAGAAGTTGGATGGAAAGCTATAGATTTCTTAATTG
AGACTGGGAATGTTAATGAACTAGAACAGTATAGGAATTATTTAAGATCACTATTGTGGCATAGATTGCA
AGGAGTTAGAGACGATGCTTGGAAACACTTACATGTGTATA

Exercise 1:

Questions:

- What is the correct open reading frame of the protein?
- Determine the start and stop codon.

First the nucleotide sequence should be translated in all possible 6 reading frames.

Go to Bioinformatics Resource Portal (ExPASy) via **http://www.expasy.org/**

ExPASy is the SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. On this portal you find resources from many different SIB groups as well as external institutions.

- Resources A…Z
- Translate

Procedure:

- Cut and paste sequence (without headings) in the entry field (Strg C/StrgV)
- TRANSLATE SEQUENCE
- Select the correct open reading frame
- Choose start position
- Export sequence in FASTA format
- Copy into a word document

Exercise 2:

Does the nucleotide sequence encode a protein? Are there any homologous with significant similarity in the data bases?
Perform a BLAST search! Consider if the similarity covers the whole protein length.

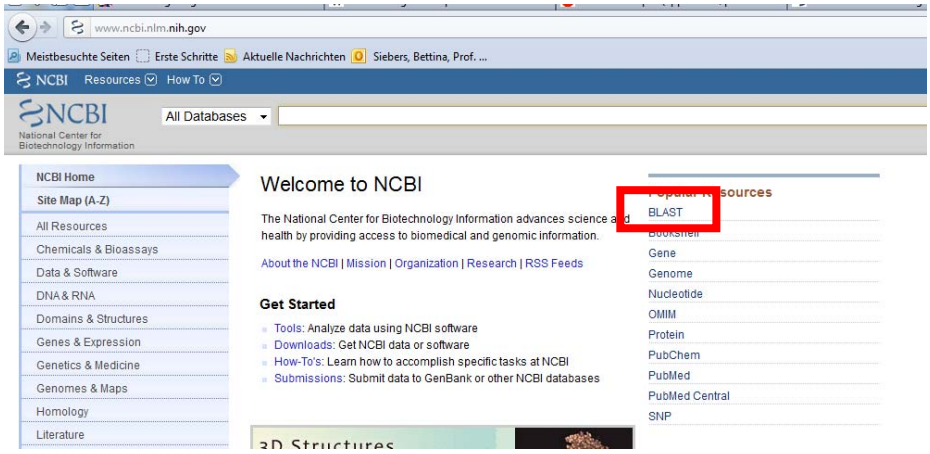
Question:

- Can you predict a function for the protein?

Procedure:

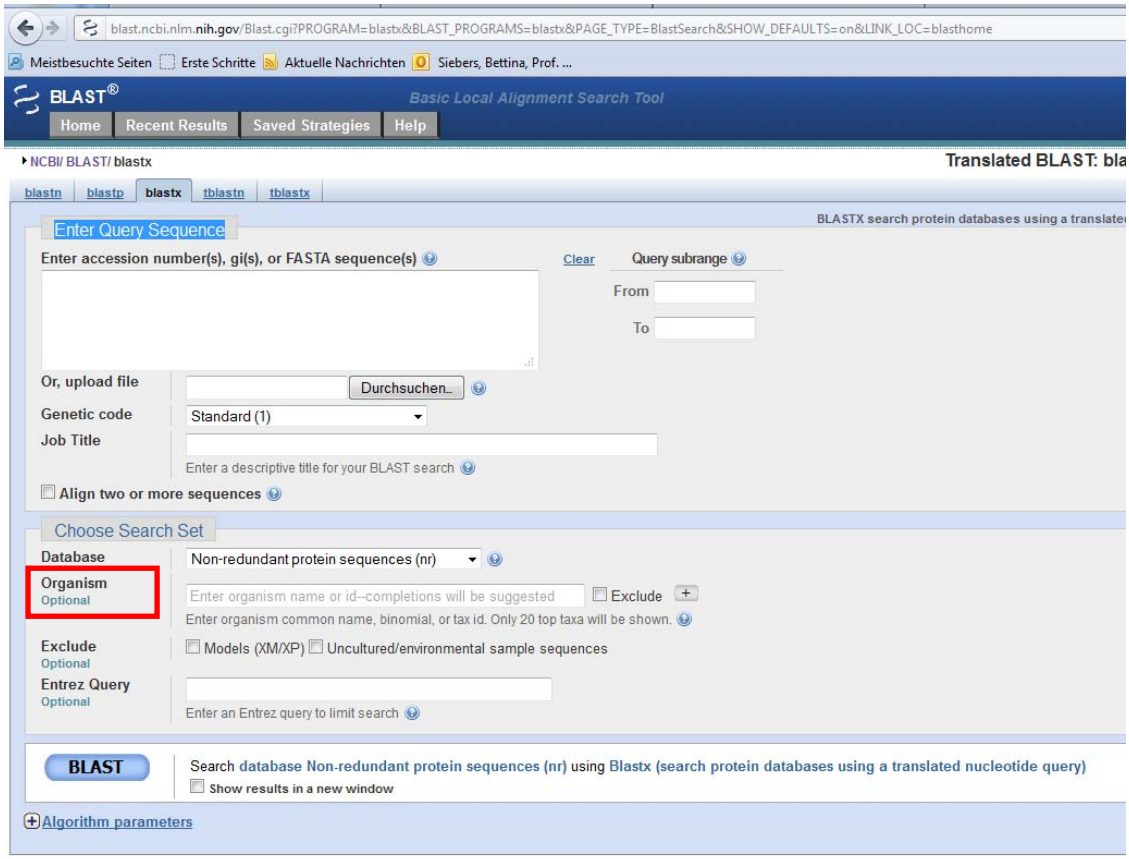
Go to the National Center for Biotechnology Information (NCBI) via

http://www.ncbi.nlm.nih.gov/

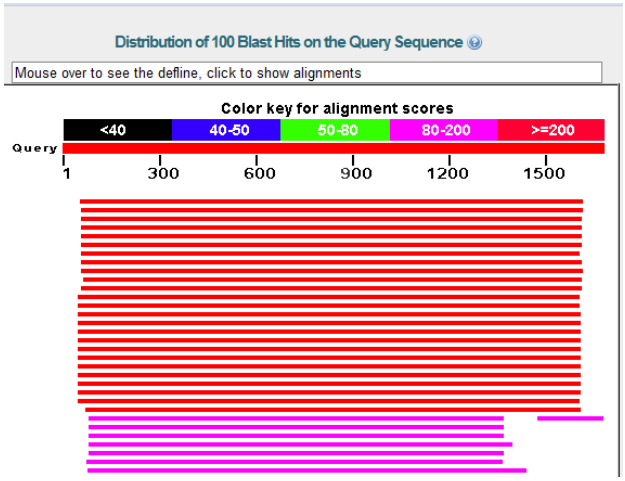


- Select:BLAST
- Select the correct Basic BLAST:
  protein blast: Search protein database using a protein query
- **Enter Query Sequence:** Copy & paste the translated protein sequence
- **Options:** Organism: specific groups (e.g. Archaea) or Organisms (e.g. *Sulfolobus solfataricus* P2) might be selected (usually the standard settings are fine)
- Use standard settings

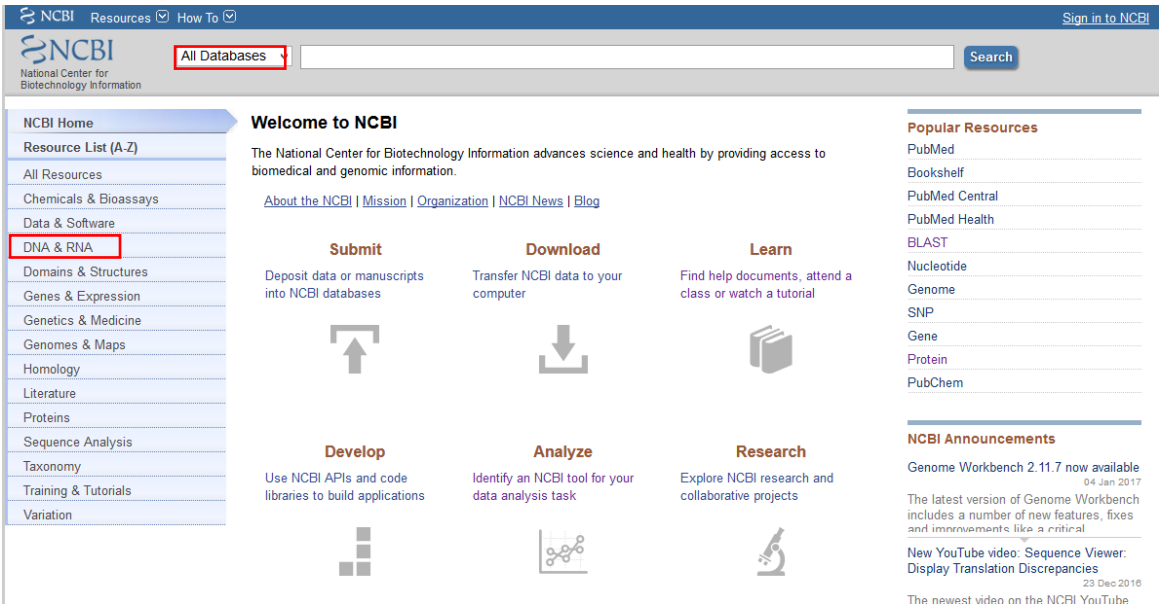
- **Blast !**



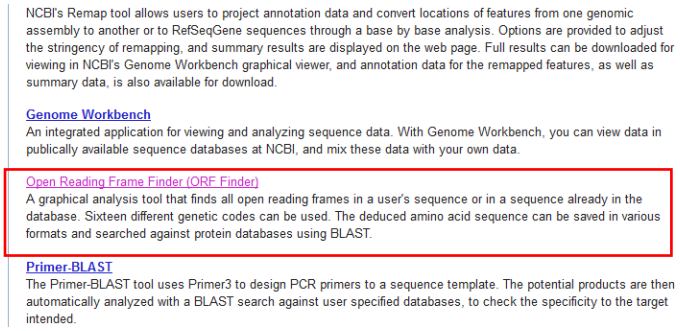
Results !



Alternatively, you may also try the “Orf Finder” at NCBI (https://www.ncbi.nlm.nih.gov/):

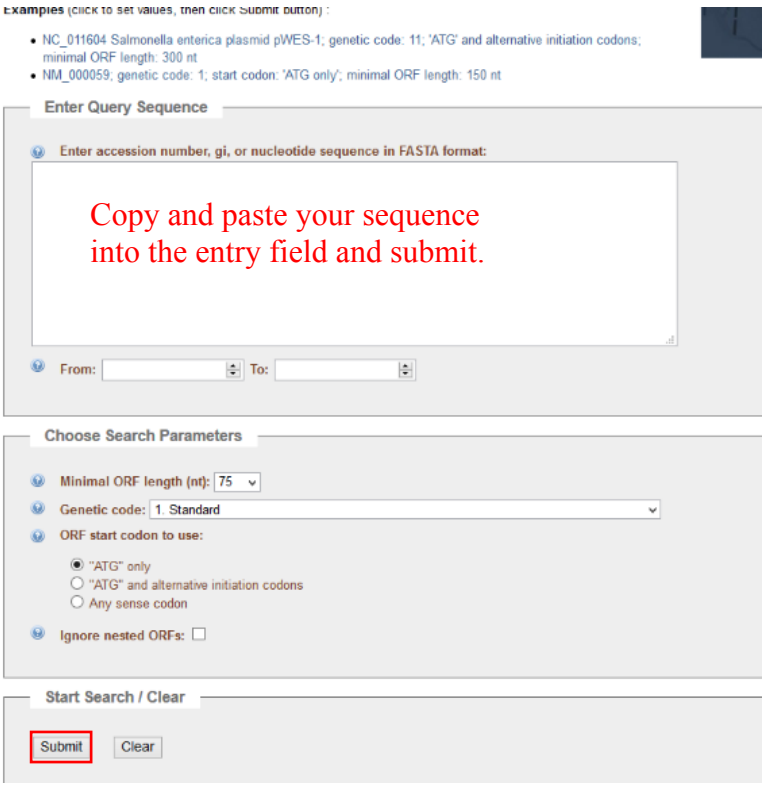


Select DNA&RNA

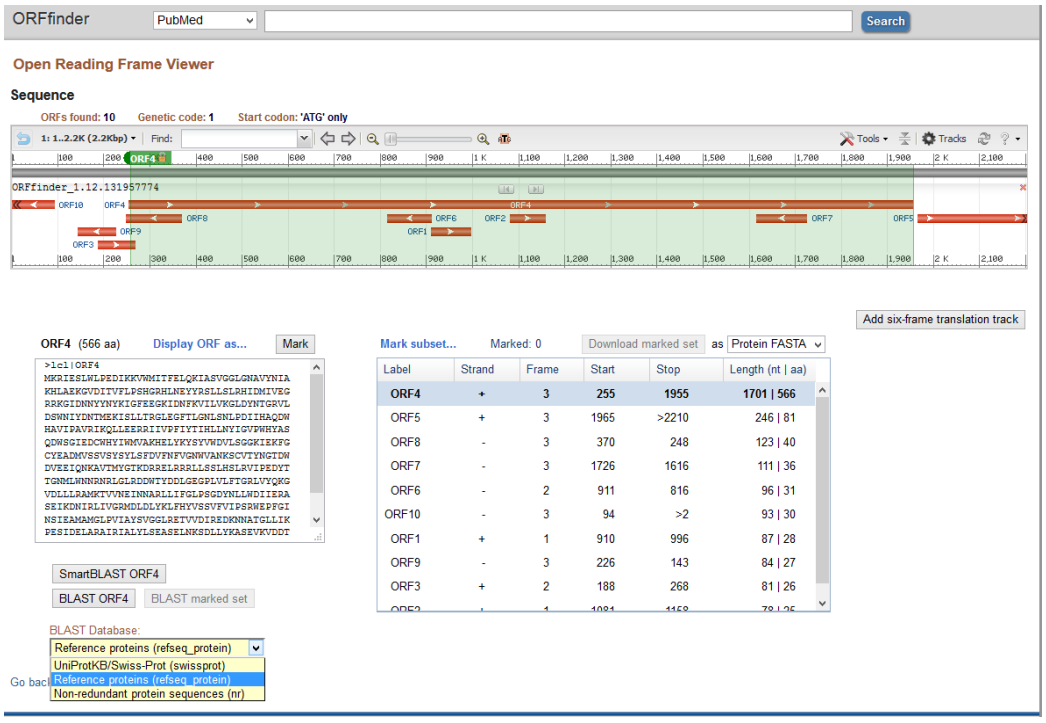


Scroll down to “Tools” and select “ORF Finder”.

(https://www.ncbi.nlm.nih.gov/orffinder/)



Inspect the results.

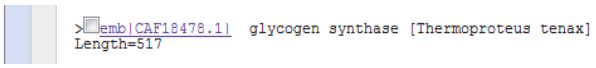


You can also directly blast the encoded protein(s), use the refseq_proteins database.

Exercise 3:

What is the enzyme activity of the gene product? In which metabolic pathway is the enzyme involved? What is known about the enzyme (biochemistry, enzymatics (e.g. effectors))?

Procedure:
For additional information click on the „Sequence ID" of a „hit"



Here you learn something about the organism, the sequence, annotation, publication(s) and if known the predicted/possible function (click on EC number of protein if present) of the enzyme.

Protein 1..517
/product="XYZ"
/EC_number="2.4.1.11"

Direct connection: ExPASy ENZYME: ECxxxx
http://expasy.org/
ExPASy Proteomics -> protein sequences and identification -> function analysis

Cross-references:
Here you find additional interesting links:

Go to:
1) –BRENDA data base
Here you find for example the enzyme reaction (reaction diagram), alternative names, detailed biochemical/enzymatic descriptions of characterized enzymes (K_m-, V_{max}-values, effectors etc.), links to the respective literature. There is also a link to the respective pathway (pathway map) via KEGG, see below.

BRENDA The Comprehensive Enzyme Information System
http://www.brenda-enzymes.org/

2) -KEGG Ligand Database for Enzyme Nomenclature (e.g.: more general enzyme information, pathway maps, Links to literature)

-KEGG http://www.genome.jp/kegg/

3) -MEDLINE (links to literature on the respective EC number)
(PubMed at NCBI: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed)

Exercise 4:

Comparative Genomics: For Bacteria and Archaea (operon structures, functional organization in the genome) it is often possible to get ideas about a possible function of a

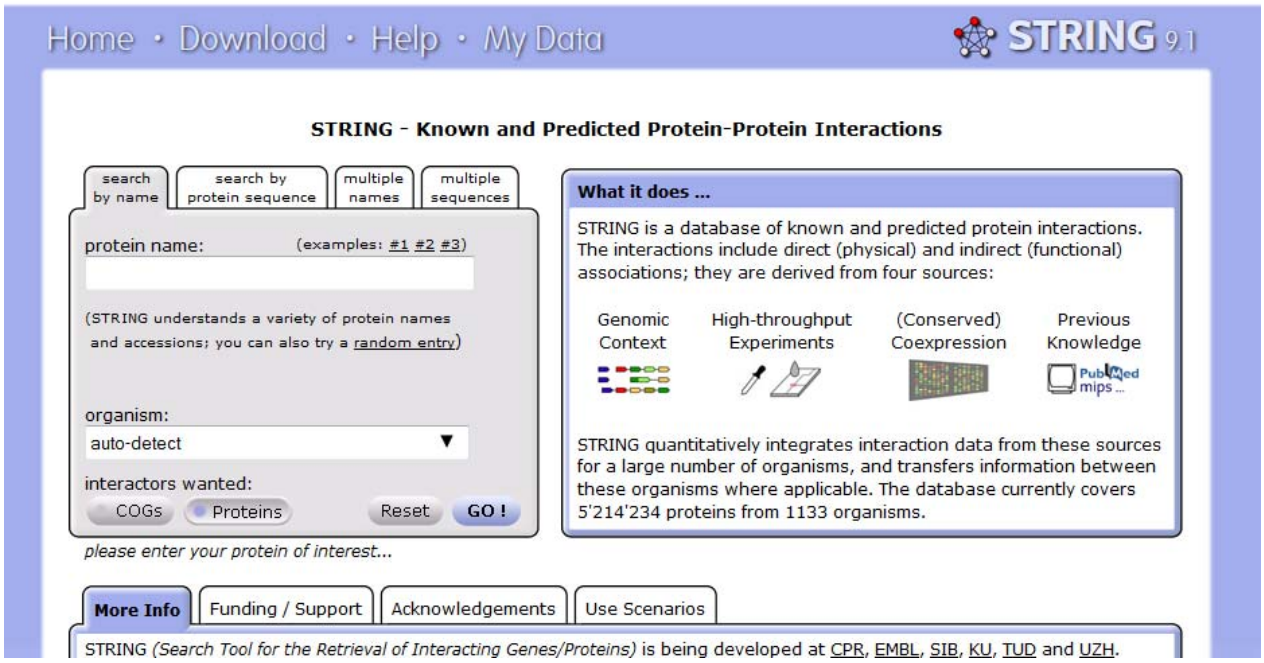
protein by analyzing the distribution in other organisms (BLAST) and the genome organization.

Questions:

- Is the encoding gene found in other organisms (e.g. Archaea)? Is there a similar gene organization?

Procedure:
Go to String 9.1 –Known and Predicted Protein-Protein Interactions

http://string-db.org/newstring_cgi/show_input_page.pl?UserId=anoYNWN0IjfB&sessionId=sAFu0BEaoaK2



- **Select: Search by protein sequence**
- **Go!**
- **Select Neighborhood**

Results !