

# 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)

```
GAGTCCAGCTGAAGTACATTCCTATTTTAAACGCCTTTGGATCCCCTATAGATGCGTTTATAAATGAATTT
TATGCAATATCGACATTTTATACATGAAGAAATAAGTAAATTAATATTAAGAATGAGCCTTATATATCA
TATTAGGAGATAAGAGAGCGTCGATAGCTTGGAAATGAAAAAGAGTTCATGGAAATTGTAATGAGAGATGA
AAGGTTTAAAACCTATTTGAAAAACTTAAGGCTGTGGTTAGGAAATGAAAAGGATTGAATCATTGTGGTT
ACCTGAGGACATTAATAAAGTGTGGATGATAACATTTGAATTGCAAAAAGATAGCAAGTGTGGAGGATTA
GGAAATGCTGTATATAACATAGCTAAACATCTAGCTGAAAAGGGAGTAGATATTACGGTTTTCTTGCCAT
CTCATGGAAGGCATCTAAATGAATATTATAGGTCTCTCTTAAGTTTAAAGACATATTGACATGATCGTTGA
GGGAAGAAGAAAGGGTATAGATAATAACTATTATAATTATAAAATAGGATTTGAAGAAGGAAAAATAGAT
AATTTCAAGGTAATTTTGGTAAAAGGATTAGATTATAACACTGGTAGGGTATTAGATTCATGGAATATCT
ATGATAACACAATGGAGAAAATTTGCTATTAACAAGAGGATTAGAGGGGTTCACTTTAGGCAATCTATC
TAACCTTCCAGATATAATCCATGCACAAGATTGGCACGCCGTAATTCGCCGAGTTAGGATAAAAACAACCTC
TTGGAAGAGAGACGAATAATCGTCCCATTTATTTACACTATTCACTTACTGAATTACATTGGTGTACCTT
GGCACTACGCTTCTCAAGACTGGTCTGGAATTGAGGATTGTTGGCATTATATTTGGATGGTAGCTAAGCA
TGAATTATATAAATATTCAATATGTATGGGATGTGTTGTGAGGTGGAAAAATTGAGAAATTTGGTTGTAC
GAGGCTGATATGGTGAGTAGTGTAAAGTTACAGTTATTTAAGTTTTGACGTATTTAATTTTTGTAGGAAAT
GGGTAGCCAATAAATCATGTGTAACGTATAATGGTACGGATTGGGATGTGGAAGAGATCCAAAAAAGGC
TGTTACAATGTATGGAAC TAAGGACAGAAGGGAGTTAAGGAGAAGGCTACTTTTCATCTCTGCACTCCCTA
AGGGTTATTCCAGAGGATTACACTACCGGCAATATGCTATGGAATAATAGAAATAGACTAGGATTAAGAG
ATGATTGGACTTACGACGATTTAGGGGAAGGACCCCTTGTCTATTCACTGGGAGATTAGTATATCAAAA
AGGTGTAGATTTACTTTTAAAGGGCTATGAAAACAGTTGTGAATGAGATAAATAATGCTAGGCTCTTAATT
TTTGGGCTACCGTCTGGGGATTACAACCTTACTTTGGGATATTATAGAGAGGGCTTCAGAAATTAAGGATA
ATATTAGGTTAATAGTGGGTAGAATGGATTTAGATTTATATAAATTGTTTCACTACGTATCTTCAGTCTT
CGTCATTCCATCTAGATGGGAACCATTTGGCATAAATTCAATTGAGGCTATGGCTATGGGGTTGCCAGTA
ATTGCCTATAGCGTAGGGGTTTTAAGAGAAAACAGTCGTTGACATTAGGGAGGATAAGAATAATGCCACAG
GTCTCTTAATTAAGCCTGAAAGTATAGACGAATTAGCTAGAGCCATAAGGATTGCATTATATTTGTCAGA
GGCTTCTGAGCTAAATAAAAGCGATTTGTTATATAAGGCGAGTGAGGTTAAAGTAGATGATACGAGATAC
TGGGATAAGGTACGTGAAAATGCAATAACTAGGGTTAAGAGTAGATTTAGATGGGATGCAGTGATAAACT
CCTTAACTGAGTGTTATAGAAAGACTCTCGATATGGCTAAATATAGGGCTTTAGCCTCTTTTTGAGGAGA
TATCATGAATATAAAGAACGCATTGGAAAGAAAGGACGTCAAATATCTGATAAGGAATATAAGAAGTTTA
CTCAACTTGCTAAAAGTAAAGATGGATTAGAAAGAGAAGTTGGATGGAAAGCTATAGATTTCTTAATTG
AGACTGGGAATGTTAATGAACTAGAACAGTATAGGAATTTAAGATCACTATTGTGGCATAGATTGCA
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.

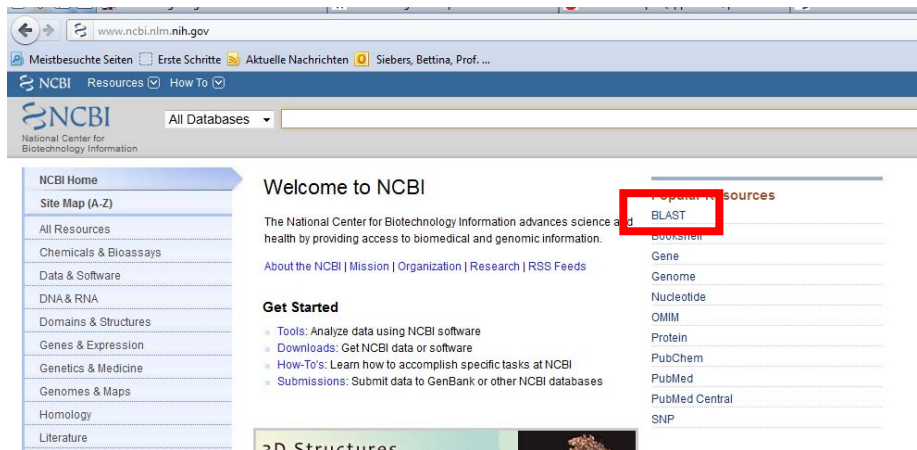
Question:

- Can you predict a function for the protein?

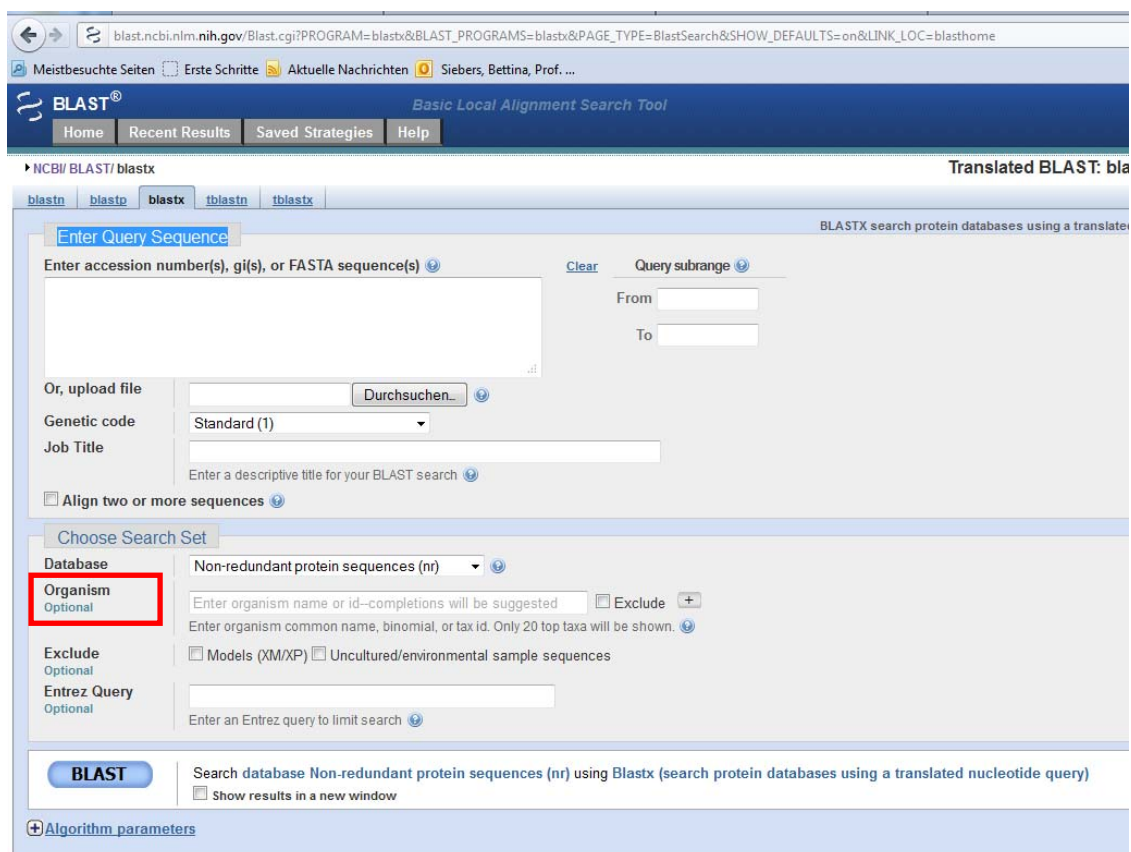
**Procedure:**

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

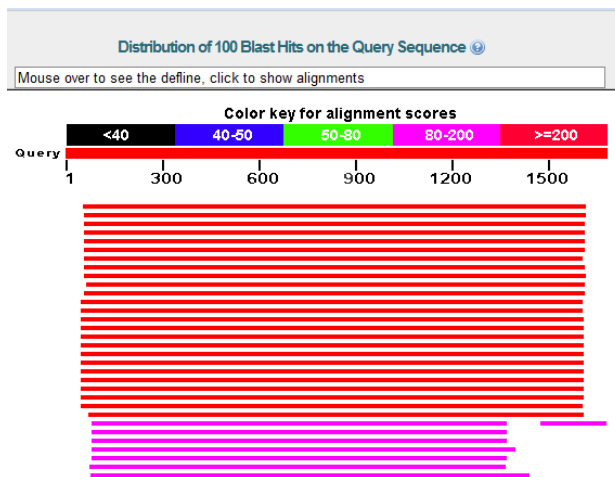
<http://www.ncbi.nlm.nih.gov/>



- 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

NCBI's Remap tool allows users to project annotation data and convert locations of features from one genomic assembly to another or to RefSeqGene sequences through a base by base analysis. Options are provided to adjust the stringency of remapping, and summary results are displayed on the web page. Full results can be downloaded for viewing in NCBI's Genome Workbench graphical viewer, and annotation data for the remapped features, as well as summary data, is also available for download.

### [Genome Workbench](#)

An integrated application for viewing and analyzing sequence data. With Genome Workbench, you can view data in publically available sequence databases at NCBI, and mix these data with your own data.

### [Open Reading Frame Finder \(ORF Finder\)](#)

A graphical analysis tool that finds all open reading frames in a user's sequence or in a sequence already in the database. Sixteen different genetic codes can be used. The deduced amino acid sequence can be saved in various formats and searched against protein databases using BLAST.

### [Primer-BLAST](#)

The Primer-BLAST tool uses Primer3 to design PCR primers to a sequence template. The potential products are then automatically analyzed with a BLAST search against user specified databases, to check the specificity to the target intended.

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

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

Examples (click to set values, then click Submit button):

- NC\_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM\_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

**Enter Query Sequence**

Enter accession number, gi, or nucleotide sequence in FASTA format:

Copy and paste your sequence into the entry field and submit.

From:  To:

**Choose Search Parameters**

Minimal ORF length (nt): 75

Genetic code: 1. Standard

ORF start codon to use:

"ATG" only

"ATG" and alternative initiation codons

Any sense codon

Ignore nested ORFs:

**Start Search / Clear**

Inspect the results.

ORFfinder PubMed

**Open Reading Frame Viewer**

Sequence

ORFs found: 10 Genetic code: 1 Start codon: 'ATG' only

1: 1..2.2K (2.2Kbp) Find:  Tools Tracks ?

ORFfinder\_1.12.131957774

**ORF4 (566 aa)** Display ORF as... Mark

```
>1c1 ORF4
MGRLESMLPEDIKKVMITFELQKIASVGGNGNAVYNIA
KHLAEKGVDTVFLPSHGRHLNEYVRSLLSLRHIDMIVGQ
RRKGINNNTYKIFGFEKIDNFFVLLVGLDINTIGVVL
DQWNYINTEHISLLPGLGFTGHSNLPDIIHAGQW
HAVIPAVRIKQLLEERLIIVPFIYIHLINVIQVPHYAS
QWNSGIEDCHHVIWAKHLYKYSVWVNDLGGKIERFG
CYEADMVSSVSYLSDFDNFVGNVANKSCVYINGIDM
DVEIQKAVYVYIKDRRELRRLLSLSLAVIPEDYT
TQMLHNDNRLLGCDWYFDLGGPFIYTORIVYKGF
VQLLDAMKIVYNEINNARLLFGLPSGVNLLWDEIEDA
SEIKDILRIVGRMDLLKLFHYVSVFVIPSRRPEFGI
NSEIAMAMGLPIAYVGGVLLRETVVDIREKNNATGLLIK
PESIDELARAIKIALYLSAESEINXSDLLYKASEVYVDDT
```

SmartBLAST ORF4

BLAST ORF4 BLAST marked set

BLAST Database:

Reference proteins (refseq\_protein)

UniProtKB/Swiss-Prot (swissprot)

Reference proteins (refseq\_protein)

Non-redundant protein sequences (nr)

Go back

Label	Strand	Frame	Start	Stop	Length (nt   aa)
ORF4	+	3	255	1955	1701   566
ORF5	+	3	1965	>2210	246   81
ORF8	-	3	370	248	123   40
ORF7	-	3	1726	1616	111   36
ORF6	-	2	911	816	96   31
ORF10	-	3	94	>2	93   30
ORF1	+	1	910	996	87   28
ORF9	-	3	226	143	84   27
ORF3	+	2	188	268	81   26
ORF2	-	1	1081	1169	79   26

Add six-frame translation track

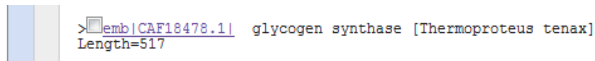
You can also directly blast the encoded protein(s), use the refseq\_proteins database.

### 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.



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=anoYNWN0ljb&sessionId=sAFu0BEaok2](http://string-db.org/newstring.cgi/show_input_page.pl?UserId=anoYNWN0ljb&sessionId=sAFu0BEaok2)

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

**Results !**