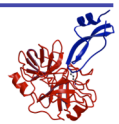
Databases

Contents, structure and annotation

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

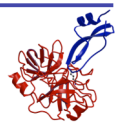


Content

- Sequence file formats
- Sequence databases
- Protein family and motif databases
- Genome and organism specific databases
- Other important databases

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway



Sequence file formats

Several sequence file formats

Contain sequence data of nucleotide (DNA) or amino acid (protein) sequences

Challenge: Sequence analysis tool/software use different file formats

Solution: Use standard formats - readable by different programs

- different programs can read many file formats

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence file formats

Several sequence file formats - most common:

Pir

```
>P1;Q3LSM1|Q3LSM1_9VIBR
Vibrio salmonicida LFI1238 Catalase  483 amino acids
MSKKLTTAAG CPVAHNQNVQ TAGKRGPQLL QDVWFLEKLA HFDREVIPER
RMHAKGSGAY GTFTVTHDIT KYTKAKIFSD IGKKTDMFAR FSTVAGERGA
ADAERDIRGF SLKFYTEEGN WDLAGNNTPV FFLRDPLKFP DLNHAVKRDP
RTNMRSAKNN WDFWTSLPEA LHQVTIVMSD RGIPATYRHM HGFGSHTFSF
INSDNERYWV KFHFVSQQGI KNLSDAEAGE LVGNDRESHQ RDLLDSIDNQ
DFPKWTLKVQ IMPEADAATV PYNPFDLTKV WPHKDYPLIE VGEFELNRNP
QNYFAEVEQA AFNPANVVPG ISFSPDKMLQ GRLFAYGDAQ RYRLGVNHQH
IPVNAPRCPV HSYHRDGAMR VDGNFGSTLG YEPNDQGQWA EQPDFSEPPL
NLDGAAAHWD HREDEDYFSQ PGDLFGLMTA EKQAILFDNT ARNLNGVPKE
IQLRHVTHCY KADPAYGEGI GKLLGFDISE YNS*
```

Fasta

```
>Q3LSM1|Q3LSM1_9VIBR Catalase - Vibrio salmonicida LFI1238.
MSKKLTTAAGCPVAHNQNVQTAGKRGPQLLQDVWFLEKLAHFDREVIPERRMHAKGSGAY
GTFTVTHDITKYTKAKIFSDIGKKTDMFARFSTVAGERGAADAERDIRGFSLKFYTEEGN
WDLAGNNTPVFFLRDPLKFPDLNHAVKRDPRTNMRSAKNNWDFWTSLPEALHQVTIVMSD
RGIPATYRHMHGFGSHTFSFINSDNERYWVKFHFVSQQGIKNLSDAEAGELVGNDRESHQ
RDLLDSIDNQDFPKWTLKVQIMPEADAATVPYNPFDLTKVWPHKDYPLIEVGEFELNRNP
QNYFAEVEQAAFNPANVVPGISFSPDKMLQGRLFAYGDAQRYRLGVNHQHIPVNAPRCPV
HSYHRDGAMRVDGNFGSTLGYEPNDQGQWAEQPDFSEPPLNLDGAAAHWDHREDEDYFSQ
PGDLFGLMTAEKQAILFDNTARNLNGVPKEIQLRHVTHCYKADPAYGEGIGKLLGFDISE
YNS
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence file formats

File converter programs

Can convert between file formats

Fasta file  **Pir file**

Readseq -http://www.ebi.ac.uk/cgi-bin/readseq.cgi

Fmtseq -http://bioinformatics.org/JaMBW/1/2/

RSA-tool -http://rsat.ulb.ac.be/rsat/convert-seq.html

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence file formats

Other important sequence file formats:

EMBL (European Molecular Biology Laboratory) format

Genbank (National Center for Biotechnology Information - NCBI) format

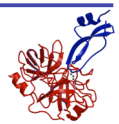
```
LOCUS       Q3LSM1|Q3LSM1_9VIBR483 aa
DEFINITION  Q3LSM1|Q3LSM1_9VIBR Catalase - Vibrio salmonicida LFI1238.
ORIGIN
        1 MSKKLTTAAG CPVAHNQNVQ TAGKRGPQLL QDVWFLEKLA HFDREVIPER RMHAKGSGAY
       61 GTFTVTHDIT KYTKAKIFSD IGKKTDMFAR FSTVAGERGA ADAERDIRGF SLKFYTEEGN
      121 WDLAGNNTPV FFLRDPLKFP DLNHAVKRDP RTNMRSAKNN WDFWTSLPEA LHQVTIVMSD
      181 RGIPATYRHM HGFGSHTFSF INSDNERYWV KFHFVSQQGI KNLSDAEAGE LVGNDRESHQ
      241 RDLLDSIDNQ DFPKWTLKVQ IMPEADAATV PYNPFDLTKV WPHKDYPLIE VGEFELNRNP
      301 QNYFAEVEQA AFNPANVVPG ISFSPDKMLQ GRLFAYGDAQ RYRLGVNHQH IPVNAPRCPV
      361 HSYHRDGAMR VDGNFGSTLG YEPNDQGQWA EQPDFSEPPL NLDGAAAHWD HREDEDYFSQ
      421 PGDLFGLMTA EKQAILFDNT ARNLNGVPKE IQLRHVTHCY KADPAYGEGI GKLLGFDISE
      481 YNS
//
```

These format are used by many sequence analysis tools - e.g. Artemis

Normally contain much more than just sequence information

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway



Sequence file formats

Other important sequence file formats:

EMBL (European Molecular Biology Laboratory) format

Genbank (National Center for Biotechnology Information - NCBI) format

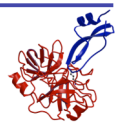
LOCUS Q3LSM1|Q3LSM1_9VIBR483 aa
DEFINITION Q3LSM1|Q3LSM1_9VIBR483 aa Catalase - Vibrio salmonicida LFI1238
ORIGIN
1 MSKKLTTAAG CPVAHNQNVQ TAGKRGPQLL QDVWFLEKLA HFDREVIPER RMHAKGSGAY
61 GTFTVTHDIT KYTKAKIFSD IGKKTDMFAR FSTVAGERGA ADAERDIRGF SLKFYTEEGN
121 WDLAGNNTPV FFLRDPLKFP DLNHAVKRDP RTNMRSAKNN WDFWTSLPEA
LHQVTIVMSD 181 RGIPATYRHM HGFGSHTFSF INSDNERYWV KFHFVSQQGI
KNLSDAEAGE LVGNDRESHQ 241 RDLLDSIDNQ DFPKWTLKVQ IMPEADAATV
PYNPFDLTKV WPHKDYPLIE VGEFELNRNP 301 QNYFAEVEQA AFNPANVVPG
ISFSPDKMLQ GRLFAYGDAQ RYRLGVNHQH IPVNAPRCPV 361 HSYHRDGAMR
VDGNFGSTLG YEPNDQGQWA EQPDFSEPPL NLDGAAAHWD HREDEDYFSQ 421
PGDLFGLMTA EKQAILFDNT ARNLNGVPKE IQLRHVTHCY KADPAYGEGI GKLLGFDISE 481
YNS

These format are used by many sequence analysis tools - e.g. Artemis

Normally contain much more than just sequence information

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway



Sequence file formats

File formats for multiple aligned sequences

MSF - Multiple Sequence Format

Header contain information about the sequences

```
  MSF:  511  Type: P    Check:  2869   ..

Name: Q3LSM1|Catalase_-_Vibrio_salmo oo  Len:  511  Check:  2193  Weight:  29.4
Name: Q7CH91|Catalase_-_Yersinia_pes oo  Len:  511  Check:   772  Weight:  33.3
Name: Q2YXT2|Catalase_-_Staphylococc oo  Len:  511  Check:  9904  Weight:  37.2
```

```
Q3LSM1|Catalase_-_Vibrio_salmo      .MS...KK.. LTTAAGCPVA HNQNVQTAGK RGPQLLQDVW FLEKLAHFDR
Q7CH91|Catalase_-_Yersinia_pes      MMS...KKKG LTTAAGAPVV DNNNVITAGR RGPMLLQDVW FLEKLAHFDR
Q2YXT2|Catalase_-_Staphylococc      .MSQQDKK.. LTGVFGHPVS DRENSMTAGP RGPLLMQDIY FLEQMSQFDR


Q3LSM1|Catalase_-_Vibrio_salmo      EVIPERRMHA KGSGAYGTFT VTHDITKYTK AKIFSDIGKK TDMFARFSTV
Q7CH91|Catalase_-_Yersinia_pes      EVIPERRMHA KGSGAYGTFT VTHDITQYTR AKIFSEIGKK TDMFVRFSTV
Q2YXT2|Catalase_-_Staphylococc      EVIPERRMHA KGSGAFGTFT VTKDITKYTN AKIFSEIGKQ TEMFARFSTV


Q3LSM1|Catalase_-_Vibrio_salmo      AGERGAADAE RDIRGFSLKF YTEEGNWDLA GNNTPVFFLR DPLKFPDLNH
Q7CH91|Catalase_-_Yersinia_pes      AGERGAADAE RDIRGFAMKF YTEEGNWDLV GNDTPVFYLR DPLKFPDLNH
Q2YXT2|Catalase_-_Staphylococc      AGERGAADAE RDIRGFALKF YTEEGNWDLV GNNTPVFFFR DPKLFVSLNR


Q3LSM1|Catalase_-_Vibrio_salmo      AVKRDPRTNM RSAKNNWDFW TSLPEALHQV TIVMSDRGIP ATYRHMHGFG
Q7CH91|Catalase_-_Yersinia_pes      VVKRDPRTNL RNPVYKWDFF SQLPEALHQL TIDFSDRGLP KSYRHMHGFG
Q2YXT2|Catalase_-_Staphylococc      AVKRDPRTNM RDAQNNWDFW TGLPEALHQV TILMSDRGIP KDLRHMHGFG


Q3LSM1|Catalase_-_Vibrio_salmo      SHTFSFINSD NERYWVKFHF VSQQGIKNLS DAEAGELVGN DRESHQRDLL
Q7CH91|Catalase_-_Yersinia_pes      SHTFSLINAN NERFWVKFHF RCQQGIENLM DDEAEKLIGS DRESSQRDLY
Q2YXT2|Catalase_-_Staphylococc      SHTYSMYNDS GERVWVKFHF RTQQGIENLT DEEAAEIIAT DRDSSQRDLF


Q3LSM1|Catalase_-_Vibrio_salmo      DSIDNQDFPK WTLKVQIMPE ADAATVPYNP FDLTKVWPHK DYPLIEVGEF
Q7CH91|Catalase_-_Yersinia_pes      EAIERGDFPR WKLQIQVMPE HEASQTPYNP FDLTKVWPHG DYPLIDVGYF
Q2YXT2|Catalase_-_Staphylococc      EAIEKGDYPK WTMYIQVMTE EQAKNHKDNP FDLTKVWYHD EYPLIEVGEF


Q3LSM1|Catalase_-_Vibrio_salmo      ELNRNPQNYF AEVEQAAFNP ANVVPGISFS PDKMLQGRLF AYGDAQRYRL
Q7CH91|Catalase_-_Yersinia_pes      ELNRNPENYF AEVEQSSFNP ANVVPGISFS PDRMLQGRLF SYGDAARYRL
Q2YXT2|Catalase_-_Staphylococc      ELNRNPDNYF MDVEQAAFAP TNIIPGLDFS PDKMLQGRLF SYGDAQRYRL


Q3LSM1|Catalase_-_Vibrio_salmo      GVNHQHIPVN APR....... CPVHSYHRDG AMR.VDGNFG STLGYEPNDQ
Q7CH91|Catalase_-_Yersinia_pes      GVNHHQIPVN SAK....... CPFHNYHRDG AMR.VDGNSG NGATYEPNSF
Q2YXT2|Catalase_-_Staphylococc      GVNHWQIPVN QPKGVGIENI CPFS...RDG QMRVVDNNQG GGTHYYPNNH


Q3LSM1|Catalase_-_Vibrio_salmo      GQWAEQPDFS EPPLNLDGAA AHWDHREDED .YFSQPGDLF GLMTAEKQAI
Q7CH91|Catalase_-_Yersinia_pes      GLFQEQPDFS EPPLTLEGAA DHWNHREDTD .YFSQPRALF NLLSEEEHQR
Q2YXT2|Catalase_-_Staphylococc      GKFDSQPEYK KPPFPTDGYG YEYNQRQDDD NYFEQPGKLF RLQSEDAKER


Q3LSM1|Catalase_-_Vibrio_salmo      LFDNTARNLN GVPKEIQLRH VTHCYKADPA YGEGIGKLLG FDIS......
Q7CH91|Catalase_-_Yersinia_pes      MFTRIAGELS QVPEAIQRRQ IALFTQVDPA YGAGVKKALG LN........
Q2YXT2|Catalase_-_Staphylococc      IFTNTANAME GVTDDVKRRH IRHCYKADPE YGKGVAKALG IDINSIDLET
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway



Sequence file formats

File formats for multiple aligned sequences

ClustalW - ALN Format

Contain no header

```
Q3LSM1|Catalase_-_Vibrio_salmo      -MS---KK--LTTAAGCPVAHNQNVQTAGKRGPQLLQDVWFLEKLAHFDR
Q7CH91|Catalase_-_Yersinia_pes      MMS---KKKGLTTAAGAPVVDNNNVITAGRRGPMLLQDVWFLEKLAHFDR
Q2YXT2|Catalase_-_Staphylococc      -MSQQDKK--LTGVFGHPVSDRENSMTAGPRGPLLMQDIYFLEQMSQFDR
                                     **   **  ** . * ** ..:*  *** *** *:**::***::::***

Q3LSM1|Catalase_-_Vibrio_salmo      EVIPERRMHAKGSGAYGTFTVTHDITKYTKAKIFSDIGKKTDMFARFSTV
Q7CH91|Catalase_-_Yersinia_pes      EVIPERRMHAKGSGAYGTFTVTHDITQYTRAKIFSEIGKKTDMFVRFSTV
Q2YXT2|Catalase_-_Staphylococc      EVIPERRMHAKGSGAFGTFTVTKDITKYTNAKIFSEIGKQTEMFARFSTV
                                    ***************:******:***:**.*****:***:*:**.*****

Q3LSM1|Catalase_-_Vibrio_salmo      AGERGAADAERDIRGFSLKFYTEEGNWDLAGNNTPVFFLRDPLKFPDLNH
Q7CH91|Catalase_-_Yersinia_pes      AGERGAADAERDIRGFAMKFYTEEGNWDLVGNDTPVFYLRDPLKFPDLNH
Q2YXT2|Catalase_-_Staphylococc      AGERGAADAERDIRGFALKFYTEEGNWDLVGNNTPVFFFRDPKLFVSLNR
                                    ****************:;***********.**:****::***  * .**:

Q3LSM1|Catalase_-_Vibrio_salmo      AVKRDPRTNMRSAKNNWDFWTSLPEALHQVTIVMSDRGIPATYRHMHGFG
Q7CH91|Catalase_-_Yersinia_pes      VVKRDPRTNLRNPVYKWDFFSQLPEALHQLTIDFSDRGLPKSYRHMHGFG
Q2YXT2|Catalase_-_Staphylococc      AVKRDPRTNMRDAQNNWDFWTGLPEALHQVTILMSDRGIPKDLRHMHGFG
                                    .********:*..  :***:: *******:** :*****:*   *******

Q3LSM1|Catalase_-_Vibrio_salmo      SHTFSFINSDNERYWVKFHFVSQQGIKNLSDAEAGELVGNDRESHQRDLL
Q7CH91|Catalase_-_Yersinia_pes      SHTFSLINANNERFWVKFHFRCQQGIENLMDDEAEKLIGSDRESSQRDLY
Q2YXT2|Catalase_-_Staphylococc      SHTYSMYNDSGERVWVKFHFRTQQGIENLTDEEAAEIIATDRDSSQRDLF
                                    ***:*: * ..** ******  ****:** * ** :::..**:* ****

Q3LSM1|Catalase_-_Vibrio_salmo      DSIDNQDFPKWTLKVQIMPEADAATVPYNPFDLTKVWPHKDYPLIEVGEF
Q7CH91|Catalase_-_Yersinia_pes      EAIERGDFPRWKLQIQVMPEHEASQTPYNPFDLTKVWPHGDYPLIDVGYF
Q2YXT2|Catalase_-_Staphylococc      EAIEKGDYPKWTMYIQVMTEEQAKNHKDNPFDLTKVWYHDEYPLIEVGEF
                                    ::*:. *:*:*.: :*:*.* :*     ********* * :****:** *

Q3LSM1|Catalase_-_Vibrio_salmo      ELNRNPQNYFAEVEQAAFNPANVVPGISFSPDKMLQGRLFAYGDAQRYRL
Q7CH91|Catalase_-_Yersinia_pes      ELNRNPENYFAEVEQSSFNPANVVPGISFSPDRMLQGRLFSYGDAARYRL
Q2YXT2|Catalase_-_Staphylococc      ELNRNPDNYFMDVEQAAFAPTNIIPGLDFSPDKMLQGRLFSYGDAQRYRL
                                    ******:*** :***::* *:*::**:.****:*******:**** ****

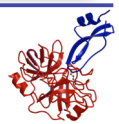
Q3LSM1|Catalase_-_Vibrio_salmo      GVNHQHIPVNAPR-------CPVHSYHRDGAMR-VDGNFGSTLGYEPNDQ
Q7CH91|Catalase_-_Yersinia_pes      GVNHHQIPVNSAK-------CPFHNYHRDGAMR-VDGNSGNGATYEPNSF
Q2YXT2|Catalase_-_Staphylococc      GVNHWQIPVNQPKGVGIENICPFS---RDGQMRVVDNNQGGGTHYYPNNH
                                    **** :**** .:       **.    *** ** **.* *.   * **.

Q3LSM1|Catalase_-_Vibrio_salmo      GQWAEQPDFSEPPLNLDGAAAHWDHREDED-YFSQPGDLFGLMTAEKQAI
Q7CH91|Catalase_-_Yersinia_pes      GLFQEQPDFSEPPLTLEGAADHWNHREDTD-YFSQPRALFNLLSEEEHQR
Q2YXT2|Catalase_-_Staphylococc      GKFDSQPEYKKPPFPTDGYGYEYNQRQDDDNYFEQPGKLFRLQSEDAKER
                                    * : .**::.:**:  :* . .:::*:* * **.**  ** * : : :

Q3LSM1|Catalase_-_Vibrio_salmo      LFDNTARNLNGVPKEIQLRHVTHCYKADPAYGEGIGKLLGFDIS------
Q7CH91|Catalase_-_Yersinia_pes      MFTRIAGELSQVPEAIQRRQIALFTQVDPAYGAGVKKALGLN--------
Q2YXT2|Catalase_-_Staphylococc      IFTNTANAMEGVTDDVKRRHIRHCYKADPEYGKGVAKALGIDINSIDLET
                                    :* . *  :. *.. :: *::    :.** ** *: * **::
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway



Sequence file formats

File formats for multiple aligned sequences

Pir and Fasta

Contain no header

```
>P1;Q3LSM1|Catalase_-_Vibrio_salmo

-MS---KK--LTTAAGCPVAHNQNVQTAGKRGPQLLQDVWFLEKLAHFDR
EVIPERRMHAKGSGAYGTFTVTHDITKYTKAKIFSDIGKKTDMFARFSTV
AGERGAADAERDIRGFSLKFYTEEGNWDLAGNNTPVFFLRDPLKFPDLNH
AVKRDPRTNMRSAKNNWDFWTSLPEALHQVTIVMSDRGIPATYRHMHGFG
SHTFSFINSDNERYWVKFHFVSQQGIKNLSDAEAGELVGNDRESHQRDLL
DSIDNQDFPKWTLKVQIMPEADAATVPYNPFDLTKVWPHKDYPLIEVGEF
ELNRNPQNYFAEVEQAAFNPANVVPGISFSPDKMLQGRLFAYGDAQRYRL
GVNHQHIPVNAPR-------CPVHSYHRDGAMR-VDGNFGSTLGYEPNDQ
GQWAEQPDFSEPPLNLDGAAAHWDHREDED-YFSQPGDLFGLMTAEKQAI
LFDNTARNLNGVPKEIQLRHVTHCYKADPAYGEGIGKLLGFDIS------
----EYNS---
*
>P1;Q7CH91|Catalase_-_Yersinia_pes

MMS---KKKGLTTAAGAPVVDNNNVITAGRRGPMLLQDVWFLEKLAHFDR
EVIPERRMHAKGSGAYGTFTVTHDITQYTRAKIFSEIGKKTDMFVRFSTV
AGERGAADAERDIRGFAMKFYTEEGNWDLVGNDTPVFYLRDPLKFPDLNH
VVKRDPRTNLRNPVYKWDFFSQLPEALHQLTIDFSDRGLPKSYRHMHGFG
SHTFSLINANNERFWVKFHFRCQQGIENLMDDEAEKLIGSDRESSQRDLY
EAIERGDFPRWKLQIQVMPEHEASQTPYNPFDLTKVWPHGDYPLIDVGYF
ELNRNPENYFAEVEQSSFNPANVVPGISFSPDRMLQGRLFSYGDAARYRL
GVNHHQIPVNSAK-------CPFHNYHRDGAMR-VDGNSGNGATYEPNSF
GLFQEQPDFSEPPLTLEGAADHWNHREDTD-YFSQPRALFNLLSEEEHQR
MFTRIAGELSQVPEAIQRRQIALFTQVDPAYGAGVKKALGLN--------
----------
*
>P1;Q2YXT2|Catalase_-_Staphylococc

-MSQQDKK--LTGVFGHPVSDRENSMTAGPRGPLLMQDIYFLEQMSQFDR
EVIPERRMHAKGSGAFGTFTVTKDITKYTNAKIFSEIGKQTEMFARFSTV
AGERGAADAERDIRGFALKFYTEEGNWDLVGNNTPVFFFRDPKLFVSLNR
AVKRDPRTNMRDAQNNWDFWTGLPEALHQVTILMSDRGIPKDLRHMHGFG
SHTYSMYNDSGERVWVKFHFRTQQGIENLTDEEAAEIIATDRDSSQRDLF
EAIEKGDYPKWTMYIQVMTEEQAKNHKDNPFDLTKVWYHDEYPLIEVGEF
ELNRNPDNYFMDVEQAAFAPTNIIPGLDFSPDKMLQGRLFSYGDAQRYRL
GVNHWQIPVNQPKGVGIENICPFS---RDGQMRVVDNNQGGGTHYYPNNH
GKFDSQPEYKKPPFPTDGYGYEYNQRQDDDNYFEQPGKLFRLQSEDAKER
IFTNTANAMEGVTDDVKRRHIRHCYKADPEYGKGVAKALGIDINSIDLET
ENDETYENFEK
*
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence file formats

File formats for structural data

Protein DataBank - PDB

Rich header with information about the structure

```
HEADER    OXIDOREDUCTASE                          17-OCT-06   2ISA
TITLE     CRYSTAL STRUCTURE OF VIBRIO SALMONICIDA CATALASE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CATALASE;
COMPND   3 CHAIN: A, B, C, D, E, F, G, H;
COMPND   4 EC: 1.11.1.6;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: VIBRIO SALMONICIDA;
SOURCE   3 ORGANISM_COMMON: BACTERIA;
SOURCE   4 STRAIN: LFI1238;
SOURCE   5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   6 EXPRESSION_SYSTEM_COMMON: BACTERIA;
SOURCE   7 EXPRESSION_SYSTEM_STRAIN: UM2;
SOURCE   8 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE   9 EXPRESSION_SYSTEM_PLASMID: PET-DEST42
KEYWDS    PEROXIDASE, HEME, OXIDOREDUCTASE, IRON, 3D-STRUCTURE,
KEYWDS   2 HYDROGEN PEROXIDE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    E.K.RIISE,M.S.LORENTZEN,R.HELLAND,A.O.SMALAS,H.K.S.LEIROS,
AUTHOR   2 N.P.WILLASSEN
REVDAT   1   23-JAN-07 2ISA    0
JRNL        AUTH   E.K.RIISE,M.S.LORENTZEN,R.HELLAND,A.O.SMALAS,
JRNL        AUTH 2 H.K.S.LEIROS,N.P.WILLASSEN
JRNL        TITL   THE FIRST STRUCTURE OF A COLD-ACTIVE CATALASE FROM
JRNL        TITL 2 VIBRIO SALMONICIDA AT 1.96A REVEALS STRUCTURAL
JRNL        TITL 3 ASPECTS OF COLD ADAPTATION
JRNL        REF    ACTA CRYSTALLOGR.,SECT.D      V.  63   135 2007
JRNL        REFN   ASTM ABCRE6  DK ISSN 0907-4449
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence file formats

File formats for structural data

Protein DataBank - PDB

Rich header with information about the structure

Contain sequence and atom coordinates

Atom no | Atom | Amino Acid residue | Residue no | X | Y | Z

```
ATOM     52  N   ALA A  10     -35.913 -42.492  21.244  1.00  7.82           N
ATOM     53  CA  ALA A  10     -37.127 -41.666  21.146  1.00  8.01           C
ATOM     54  C   ALA A  10     -37.150 -40.506  22.152  1.00  8.02           C
ATOM     55  O   ALA A  10     -38.128 -39.765  22.227  1.00  8.19           O
ATOM     56  CB  ALA A  10     -38.372 -42.521  21.283  1.00  7.98           C
ATOM     57  N   GLY A  11     -36.090 -40.381  22.948  1.00  8.04           N
ATOM     58  CA  GLY A  11     -35.918 -39.239  23.838  1.00  8.59           C
ATOM     59  C   GLY A  11     -36.485 -39.407  25.234  1.00  9.24           C
ATOM     60  O   GLY A  11     -36.584 -38.425  25.968  1.00  9.28           O
ATOM     61  N   CYS A  12     -36.859 -40.635  25.604  1.00  9.71           N
ATOM     62  CA  CYS A  12     -37.368 -40.939  26.949  1.00  9.95           C
ATOM     63  C   CYS A  12     -36.203 -41.126  27.920  1.00  9.86           C
ATOM     64  O   CYS A  12     -35.239 -41.813  27.582  1.00  9.56           O
ATOM     65  CB  CYS A  12     -38.228 -42.212  26.926  1.00  9.58           C
ATOM     66  SG  CYS A  12     -38.923 -42.745  28.524  1.00 12.55           S
```

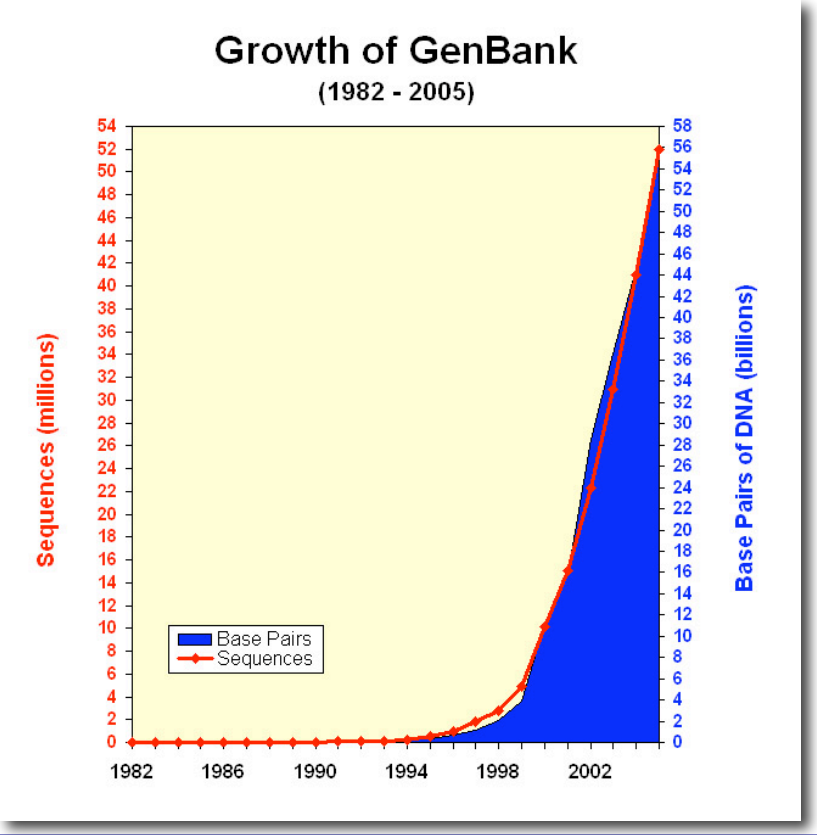
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Nucleotide

Why do we need sequence databases?

- There are over 50 million sequences in Genbank



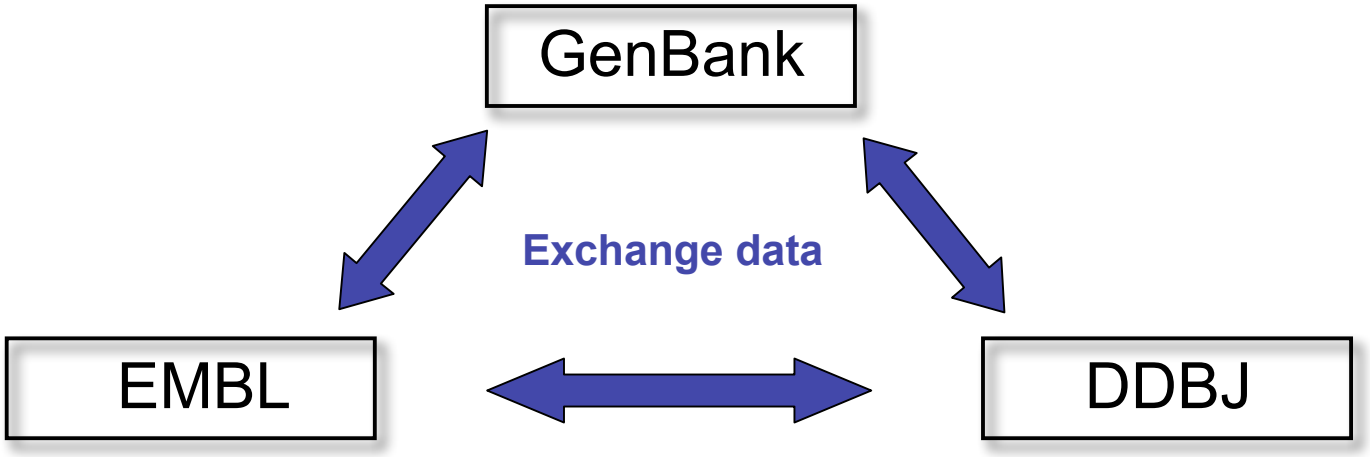
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Nucleotide

Three major nucleotide sequence databases

- Genbank, USA (http://www.ncbi.nlm.nih.gov/Genbank/index.html)
- EMBL, UK (http://www.ebi.ac.uk/embl/index.html)
- DDBJ, Japan (http://www.ddbj.nig.ac.jp/)



Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

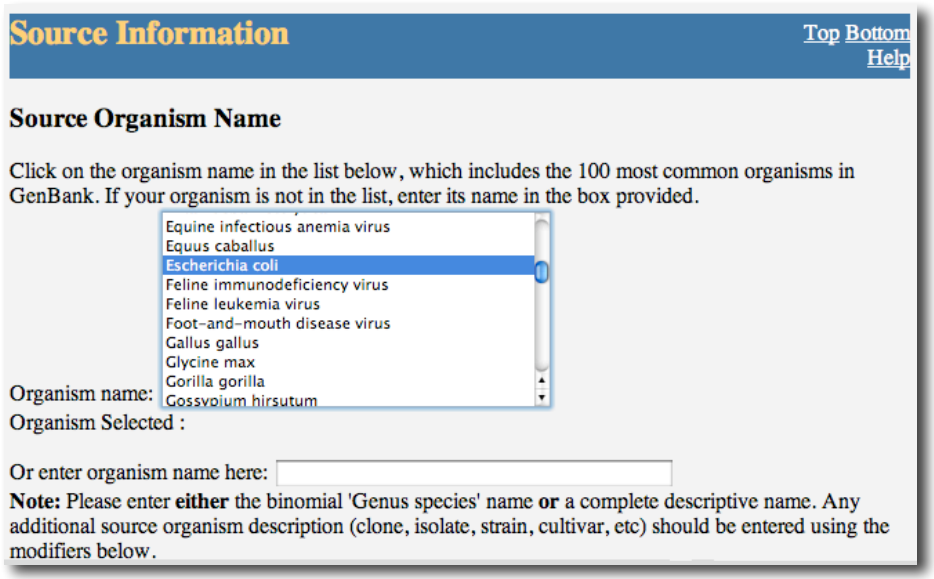
Sequence databases - Nucleotide

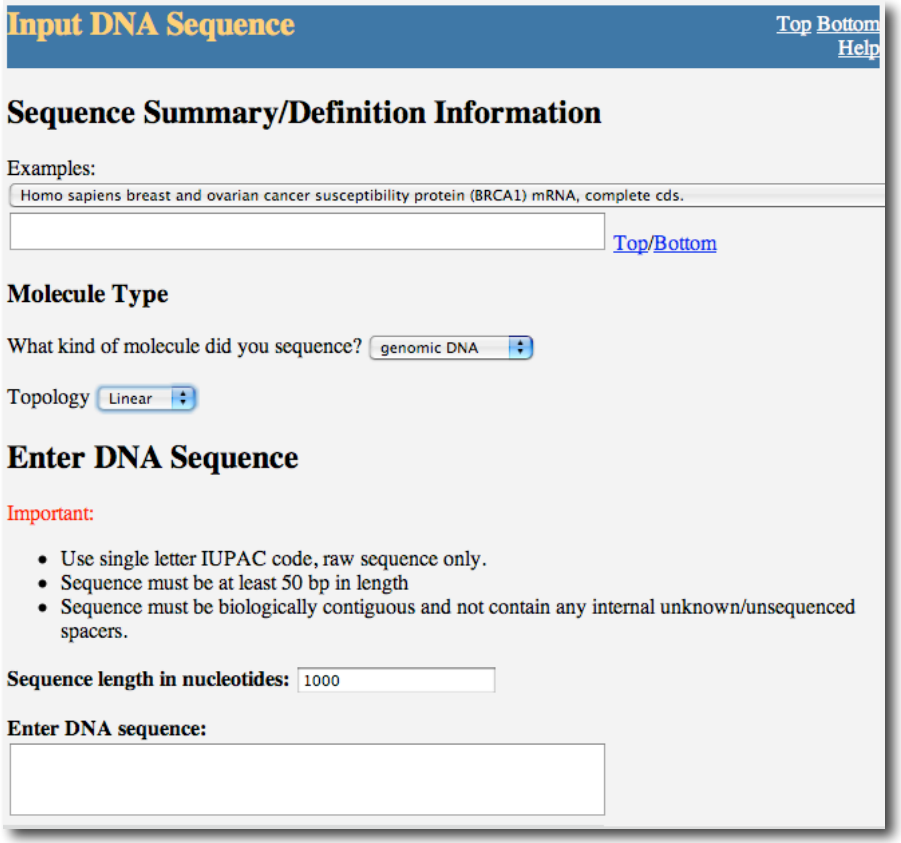
Submitting sequences

For few sequences and little complex annotations:

BankIt (GenBank)

WebIn (EMBL)





Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Nucleotide

GenBank sequence file format

Contain extensive annotation

- -Accession number
- -Name of organism and geneproduct
- -Publication, date and name of authors

```
LOCUS       DQ182487                1852 bp    DNA     linear   BCT 12-OCT-2006
DEFINITION  Vibrio salmonicida LFI1238 phospholipase gene, partial cds;
            catalase (catA) gene, complete cds; and hypothetical protein gene,
            partial cds.
ACCESSION   DQ182487
VERSION     DQ182487.1  GI:75709233
KEYWORDS    .
SOURCE      Vibrio salmonicida LFI1238
  ORGANISM  Vibrio salmonicida LFI1238
            Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales;
            Vibrionaceae; Vibrio.
REFERENCE   1  (bases 1 to 1852)
  AUTHORS   Lorentzen,M.S., Moe,E., Jouve,H.M. and Willassen,N.P.
  TITLE     Cold adapted features of Vibrio salmonicida catalase:
            characterisation and comparison to the mesophilic counterpart from
            Proteus mirabilis
  JOURNAL   Extremophiles 10 (5), 427-440 (2006)
   PUBMED   16609813
REFERENCE   2  (bases 1 to 1852)
  AUTHORS   Lorentzen,M.S. and Willassen,N.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-AUG-2005) Molecular Biotechnology, University of
            Tromso, Tromso 9037, Norway
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Nucleotide

GenBank sequence file format

Contain extensive annotation -Sequence in GenBank format

```
ORIGIN
        1 catttgtcca gccgtgatgg tactttcggg ttggacagtg aaaattgtca cttttgtatt
       61 tagcaccacc tttattaatc gctttactgt gtacttttat aataatacta aggaatgagt
      121 aatgagtaag aaattaacga cagcagcagg ttgccctgtt gcacataacc aaaatgttca
      181 aactgcgggt aaacgtggtc ctcaacttct gcaagatgtt tggtttttag aaaaactggc
      241 gcattttgat cgtgaagtta ttcctgagcg tcgtatgcac gcaaaaggtt caggtgctta
      301 tggtacgttt acggttactc atgacattac caaatacaca aaagcaaaaa tattttcaga
      361 cattggcaaa aaaacagaca tgtttgcgcg tttctctacc gttgcgggtg aacgtggcgc
      421 ggcagatgct gagcgtgata tccgtggttt ctctctgaag ttttacactg aagaaggtaa
      481 ctgggattta gcgggtaaca ataccccggt gtttttctta cgtgatccat tgaaattccc
      541 tgatttaaac cacgcagtaa aacgtgatcc tcgcacaaat atgcgaagcg cgaaaaataa
      601 ctgggatttc tggacgtcat taccagaagc actgcaccaa gtaactatcg tgatgagtga
      661 ccgtggtatt cctgcgacat accgacatat gcatggtttt ggtagccaca cattcagctt
      721 tatcaatagc gataacgaac gttattgggt gaaattccac tttgtctctc aacaaggcat
      781 taaaaaccta tcagatgcag aagcgggcga attagtcggt aatgatcgtg aaagtcatca
      841 acgtgacttg ctagacagta ttgataacca agacttccct aaatggacat tgaaagtaca
      901 aatcatgcca gaagcagatg cggcaacagt accttacaac ccatttgatt taaccaaagt
      961 atggccacac aaagattacc cattaattga agtgggtgaa tttgagctaa accgcaatcc
     1021 acaaaactat ttcgcagaag tagaacaagc ggcgtttaac cctgcaaatg tggttccggg
     1081 tatcagcttc tctcctgata aaatgctaca aggtcgtttg tttgcgtatg gtgatgccca
     1141 gcgttaccgt ttaggcgtta accaccagca catcccagtg aacgcacctc gttgtcctgt
     1201 acatagttac caccgtgatg gcgcaatgcg tgttgatggc aacttcggta gcacattagg
     1261 ttatgaacca aatgaccaag gtcaatgggc agaacaacca gatttctctg agccaccact
     1321 aaatttagat ggcgcagcag cacattggga tcaccgtgaa gatgaagatt acttctcaca
     1381 accgggtgac ttgttcggat taatgaccgc tgaaaaacaa gcaattttgt ttgataacac
     1441 agcacgtaac ctaaatggtg tgccaaaaga gattcaacta cgtcacgtaa ctcattgtta
     1501 taaagctgat ccagcgtacg gtgaaggtat tggtaaacta cttggtttcg atatcagtga
     1561 atataattcg taattaatcg cgaattataa taagtaaata tgcttaaggc agatgacgat
     1621 aatcgttgtc tgccttttt gtttaaatat acagagactt aaaaaccaca aaaactgtcc
     1681 gtatttgtat cgatagacct ccgctatact ccccgcatta ttataagccg ctaaggcaaa
     1741 tactcattgc tctggattta cgatccagcc tcgcaaaaaa ggaaagttac ccatgtctca
     1801 atatatcgtt tgtgcgttat acaaatttgt ttcattagaa gattacaccg aa
//
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Nucleotide

GenBank sequence file format

Contain extensive annotation -Size of DNA fragment and translated protein

-Location of CDS

```
source          1..1852
                /organism="Vibrio salmonicida LFI1238"
                /mol_type="genomic DNA"
                /strain="LFI1238"
                /db_xref="taxon:316275"
CDS             <1..63
                /codon_start=1
                /transl_table=11
                /product="phospholipase"
                /protein_id="ABA26663.1"
                /db_xref="GI:75709235"
                /translation="HLSSRDGTFGLDSENCHFCI"
gene            122..1573
                /gene="catA"
CDS             122..1573
                /gene="catA"
                /EC_number="1.11.1.6"
                /function="degradation of hydrogen peroxide"
                /codon_start=1
                /transl_table=11
                /product="catalase"
                /protein_id="ABA26662.1"
                /db_xref="GI:75709234"
                /translation="MSKKLTTAAGCPVAHNQNVQTAGKRGPQLLQDVWFLEKLAHFDR
                EVIPERRMHAKGSGAYGTFTVTHDITKYTKAKIFSDIGKKTDMFARFSTVAGERGAAD
                AERDIRGFSLKFYTEEGNWDLAGNNTPVFFLRDPLKFPDLNHAVKRDPRTNMRSAKNN
                WDFWTSLPEALHQVTIVMSDRGIPATYRHMHGFGSHTFSFINSDNERYWVKFHFVSQQ
                GIKNLSDAEAGELVGNDRESHQRDLLDSIDNQDFPKWTLKVQIMPEADAATVPYNPFD
                LTKVWPHKDYPLIEVGEFELNRNPQNYFAEVEQAAFNPANVVPGISFSPDKMLQGRLF
                AYGDAQRYRLGVNHQHIPVNAPRCPVHSYHRDGAMRVDGNFGSTLGYEPNDQGQWAEQ
                PDFSEPPLNLDGAAAHWDHREDEDYFSQPGDLFGLMTAEKQAILFDNTARNLNGVPKE
                IQLRHVTHCYKADPAYGEGIGKLLGFDISEYNS"
CDS             1793..>1852
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Nucleotide

GenBank vs EMBL sequence file format

EMBL uses a two-letter code to describe linecontent

```
LOCUS       DQ182487                1852 bp    DNA     linear   BCT 12-OCT-2006
DEFINITION  Vibrio salmonicida LFI1238 phospholipase gene, partial cds;
            catalase (catA) gene, complete cds; and hypothetical protein gene,
            partial cds.
ACCESSION   DQ182487
VERSION     DQ182487.1  GI:75709233
KEYWORDS    .
SOURCE      Vibrio salmonicida LFI1238
  ORGANISM  Vibrio salmonicida LFI1238
            Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales;
            Vibrionaceae; Vibrio.
REFERENCE   1  (bases 1 to 1852)
  AUTHORS   Lorentzen,M.S., Moe,E., Jouve,H.M. and Willassen,N.P.
  TITLE     Cold adapted features of Vibrio salmonicida catalase:
            characterisation and comparison to the mesophilic counterpart from
            Proteus mirabilis
  JOURNAL   Extremophiles 10 (5), 427-440 (2006)
   PUBMED   16609813
REFERENCE   2  (bases 1 to 1852)
  AUTHORS   Lorentzen,M.S. and Willassen,N.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-AUG-2005) Molecular Biotechnology, University of
            Tromso, Tromso 9037, Norway
```

```
ID   DQ182487; SV 1; linear; genomic DNA; STD; PRO; 1852 BP.
XX
AC   DQ182487;
XX
DT   21-SEP-2005 (Rel. 85, Created)
DT   14-OCT-2006 (Rel. 89, Last updated, Version 2)
XX
DE   Vibrio salmonicida LFI1238 phospholipase gene, partial cds; catalase (catA)
DE   gene, complete cds; and hypothetical protein gene, partial cds.
XX
KW   .
XX
OS   Vibrio salmonicida LFI1238
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales; Vibrionaceae;
OC   Vibrio.
XX
RN   [1]
RP   1-1852
RX   DOI; 10.1007/s00792-006-0518-z
RX   PUBMED; 16609813.
RA   Lorentzen M.S., Moe E., Jouve H.M., Willassen N.P.;
RT   "Cold adapted features of Vibrio salmonicida catalase: characterisation and
RT   comparison to the mesophilic counterpart from Proteus mirabilis";
RL   Extremophiles 10(5):427-440(2006).
XX
RN   [2]
RP   1-1852
RA   Lorentzen M.S., Willassen N.P.;
RT   ;
RL   Submitted (26-AUG-2005) to the EMBL/GenBank/DDBJ databases.
RL   Molecular Biotechnology, University of Tromso, Tromso 9037, Norway
```

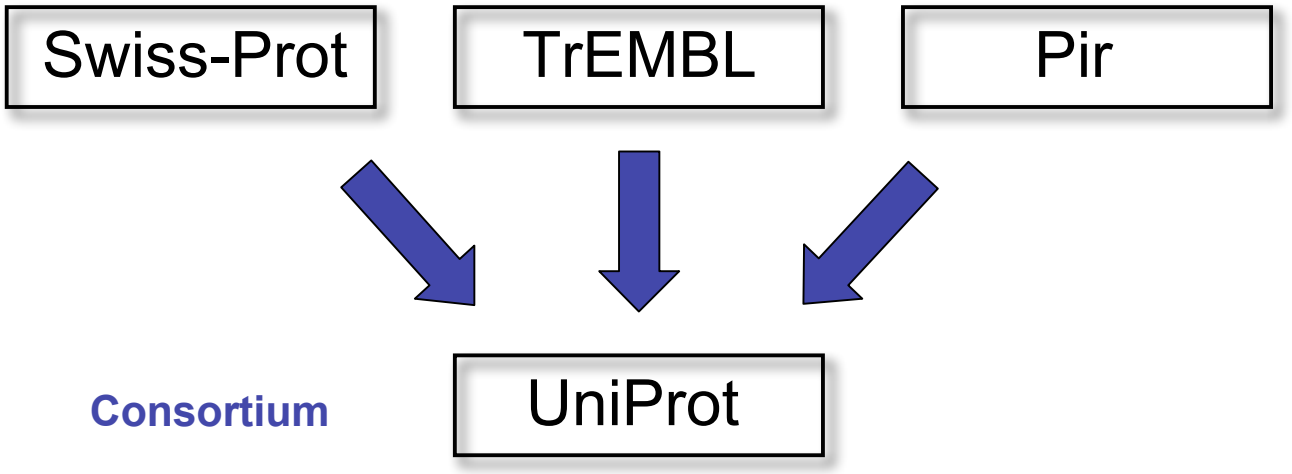
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Protein

Three protein sequence databases - one consortium

Swiss-Prot, UK and Switzerland	(http://au.expasy.org/sprot/)
TrEMBL, UK	(http://www.ebi.ac.uk/trembl/index.html)
Pir	(http://pir.georgetown.edu/pirwww/dbinfo/)
UniProt, UK, US and Switzerland	(http://www.expasy.uniprot.org/)



Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Protein

Swiss-Prot entry - manually curated entries

Entry information	
Entry name	**Q3LSM1_9VIBR**
Primary accession number	**Q3LSM1**
Secondary accession numbers	None
Integrated into TrEMBL on	October 25, 2005
Sequence was last modified on	October 25, 2005 (Sequence version 1)
Annotations were last modified on	July 24, 2007 (Entry version 16)
Name and origin of the protein	
Protein name	**Catalase**
Synonym	**EC** 1.11.1.6
Gene name	**Name: catA**
From	Vibrio salmonicida LFI1238 [TaxID: 316275]
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio.
Protein existence	3: Inferred from homology;

References

[1] NUCLEOTIDE SEQUENCE.
STRAIN=LFI1238;
DOI=10.1007/s00792-006-0518-z; PubMed=16609813 [NCBI, ExPASy, EBI, Israel, Japan]
Lorentzen M.S., Moe E., Jouve H.M., Willassen N.P.;
"Cold adapted features of Vibrio salmonicida catalase: characterisation and comparison to the mesophilic counterpart from Proteus mirabilis.";
Extremophiles 10:427-440(2006).

Comments

- ***FUNCTION***: Decomposes hydrogen peroxide into water and oxygen; serves to protect cells from the toxic effects of hydrogen peroxide *(By similarity)*.
- ***CATALYTIC ACTIVITY***: 2 H_2O_2 = O_2 + 2 H_2O.
- ***SIMILARITY***: Belongs to the catalase family.

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Sequence databases - Protein

Swiss-Prot entry - links to other resources

Cross-references	
Sequence databases	
EMBL	DQ182487; ABA26662.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence]
3D structure databases	
SMR	Q3LSM1; 3-474.
ModBase	Q3LSM1.
Organism-specific databases	
HOGENOM	[Family / Alignment / Tree]
Ontologies	
GO	GO:0004096; Molecular function: catalase activity *(inferred from electronic annotation from InterPro).* GO:0006118; Biological process: electron transport *(inferred from electronic annotation from InterPro).* GO:0006979; Biological process: response to oxidative stress *(inferred from electronic annotation from InterPro).* QuickGo view.
Family and domain databases	
InterPro	IPR002226; Catalase. IPR011614; Catalase_N. Graphical view of domain structure.
Gene3D	G3DSA:2.40.180.10; Catalase_N; 1.
PANTHER	PTHR11465; Catalase; 1.
Pfam	PF00199; Catalase; 1. Pfam graphical view of domain structure.
PRINTS	PR00067; CATALASE.
ProDom	PD000510; Catalase; 1. [Domain structure / List of seq. sharing at least 1 domain]
PROSITE	PS00437; CATALASE_1; 1. PS00438; CATALASE_2; 1.
Other	
UniRef	View cluster of proteins with at least 50% / 90% / 100% identity.

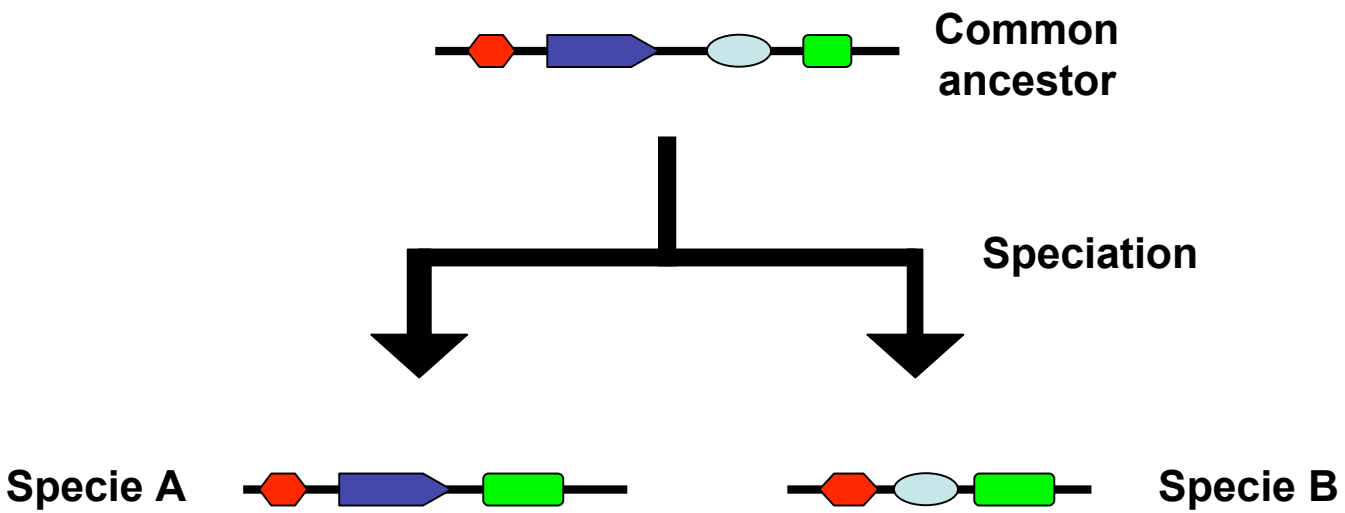
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Protein family and motif databases

Conservation of domains during evolution - can make up a protein family

e.g. DNA binding domain of a DNA binding protein



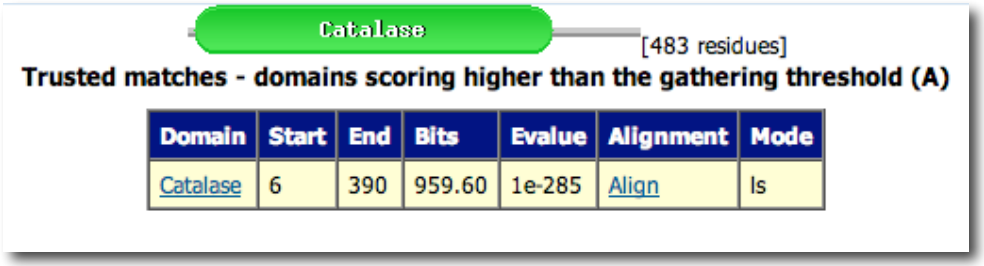
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Protein family and motif databases

Protein families - Pfam (http://www.sanger.ac.uk/Software/Pfam/)

- Database of nearly 10 000 protein families
- One protein may have more than one domain and belong to more than one family
- Contain multiple alignments of proteins belonging to the different families




```
                    *->TtnqGaPvadnqnSlTaGprGPlLLqDyhLlEklaHFdRErIPERvV
                       Tt++G+Pva nqn +TaG rGP LLqD+++lEklaHFdRE+IPER++
UNKNOWN-QU      6    TTAAGCPVAHNQNVQTAGKRGPQLLQDVWFLEKLAHFDREVIPERRM 52

                    HAkGAGAfGyFeVThDitkyTkAkflsevGkkTpvfvRFSTVaGerGSAD
                    HAkG+GA+G+F+VThDitkyTkAk++s++GkkT +f+RFSTVaGerG+AD
UNKNOWN-QU     53 HAKGSGAYGTFTVTHDITKYTKAKIFSDIGKKTDMFARFSTVAGERGAAD 102

                    TvRDpRGFAvKFYTEEGNyDLVGNNtPVFFiRDpiKFPDfIHaqKrnPrt
                    ++RD+RGF++KFYTEEGN+DL GNNtPVFF+RDp+KFPD++Ha+Kr+Prt
UNKNOWN-QU    103 AERDIRGFSLKFYTEEGNWDLAGNNTPVFFLRDPLKFPDLNHAVKRDPRT 152
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Protein family and motif databases

Protein families - Pfam (http://www.sanger.ac.uk/Software/Pfam/)

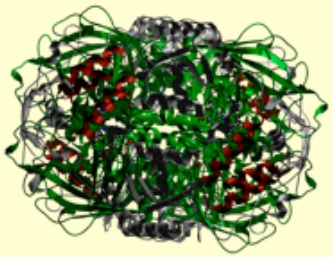


Figure 1: 1ggk
OXIDOREDUCTASE
CRYSTAL STRUCTURE OF CATALASE HPII FROM ESCHERICHIA COLI, ASN201HIS VARIANT.

Key:

Domain	Chain	Start Residue	End Residue
Catalase-rel	A	514	567
Catalase	A	81	469
Catalase-rel	B	514	567

***Accession number:* PF00199**

Previous identifiers: catalase;

Catalase Add Annotation

INTERPRO description (entry IPR011614)

Catalases () are antioxidant enzymes that catalyse the conversion of hydrogen peroxide to water and molecular oxygen, serving to protect cells from its toxic effects PUBMED:11351128. Hydrogen peroxide is produced as a consequence of oxidative cellular metabolism and can be converted to the highly reactive hydroxyl radical via transition metals, this radical being able to damage a wide variety of molecules within a cell, leading to oxidative stress and cell death. Catalases act to neutralise hydrogen peroxide toxicity, and are produced by all aerobic organisms ranging from bacteria to man. Most catalases are mono-functional, haem-containing enzymes, although there are also bifunctional haem-containing peroxidase/catalases () that are closely related to plant peroxidases, and non-haem, manganese-containing catalases () that are found in bacteria PUBMED:14745498.

This entry represents a conserved region within catalase enzymes ().

QuickGO	
Function	catalase activity (GO:0004096)
Process	response to oxidative stress (GO:0006979) electron transport (GO:0006118)

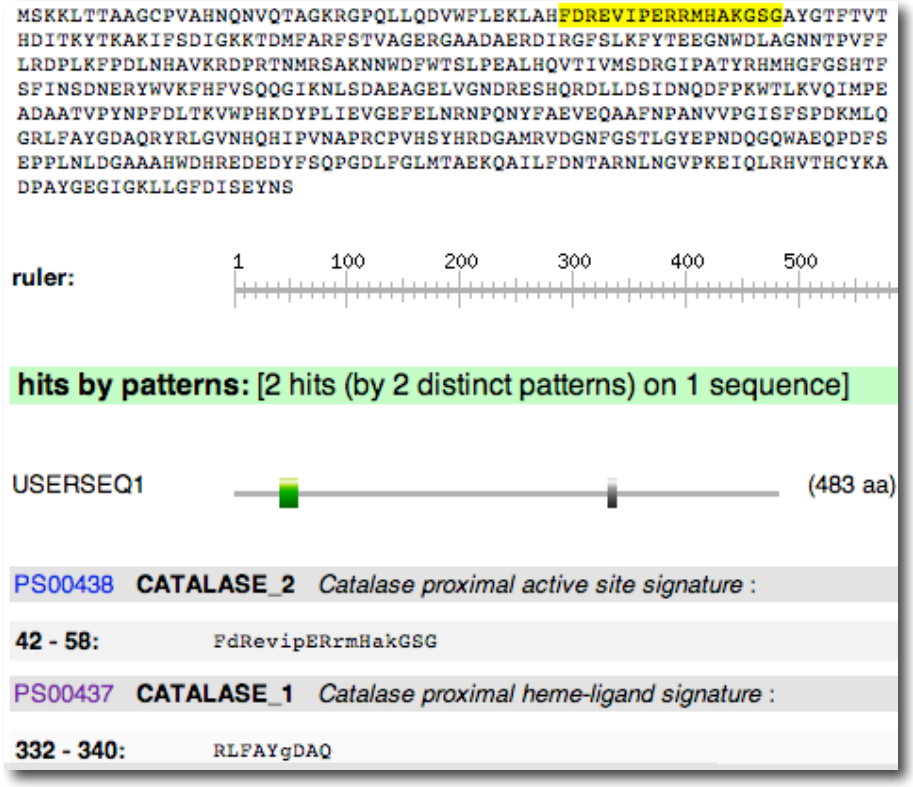
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Protein family and motif databases

Protein families - Prosite (http://au.expasy.org/prosite/)

Database of protein domains, families and functional sites



Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

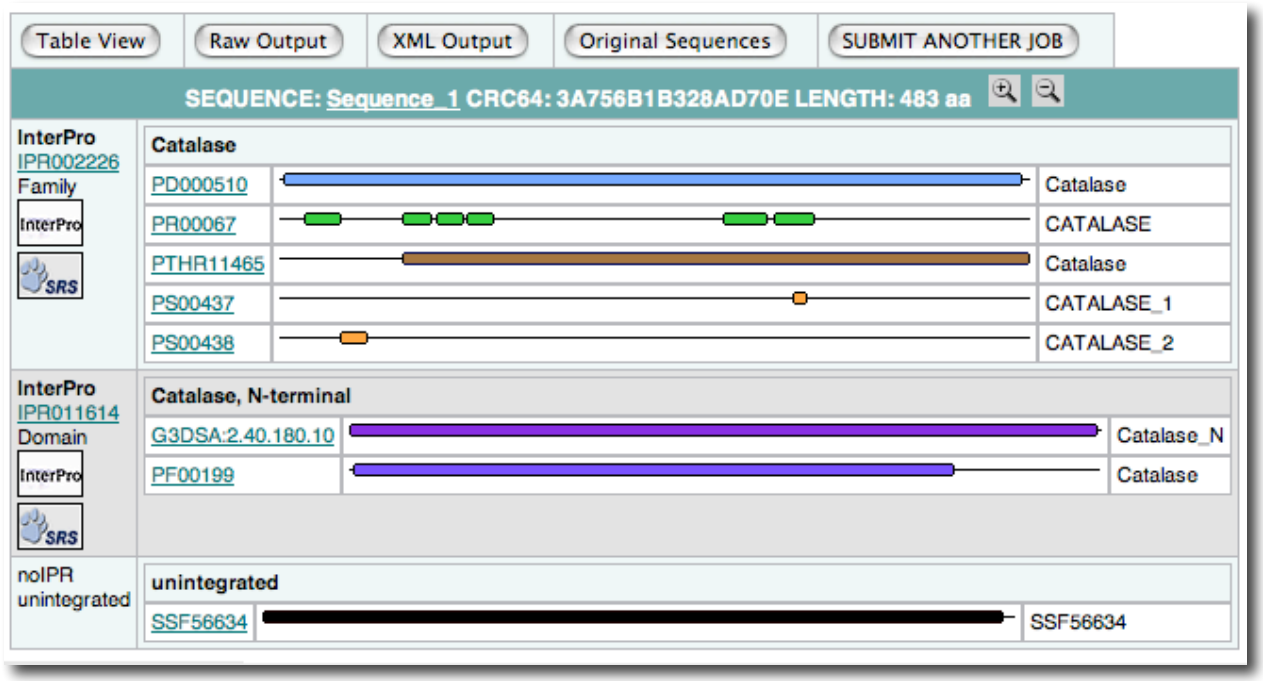
Protein family and motif databases

Protein families - InterPro (http://www.ebi.ac.uk/interpro/)

Integrated database of protein domains, families and functional sites

Contains both Pfam, Prosite and several other databases





Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

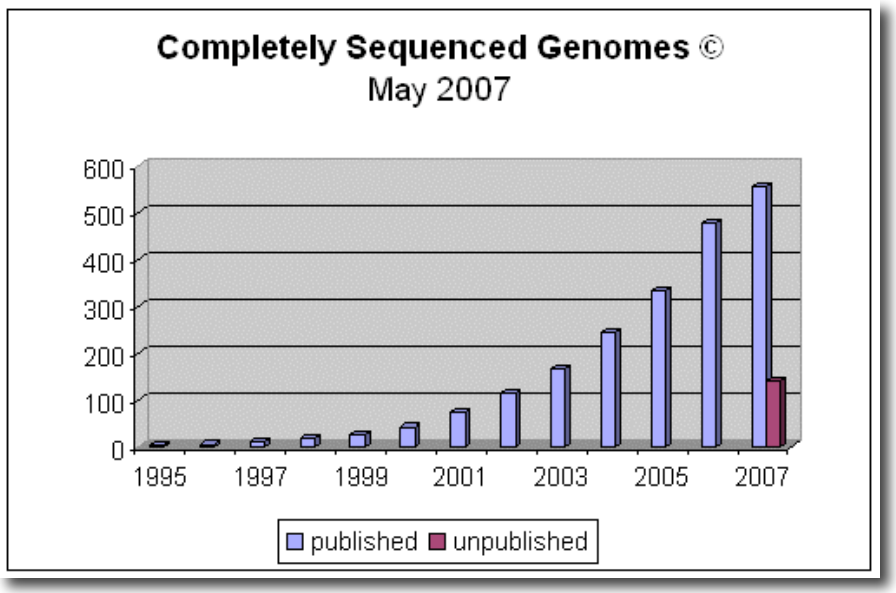
Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases

Genome databases - Complete and ongoing genome sequencing projects

1977 - Bacteriophage fx174 (5 268 bp)

1995 - Haemophilus influenzae (1.8 Mb)



2007: 634 complete genome sequences

2 229 ongoing genome sequencing projects

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases



Genome databases - Genomes Online Database (GOLD)

- Published genomes - Archaeal, bacterial and eukaryotic
- Ongoing genome projects - Archaeal, bacterial and eukaryotic
- Metagenome projects

Gc00113	*Escherichia coli O6:K2:H1* UPEC-CFT073	B	PROTEOBACTERIA-GAMMA Taxonomy Entrez Wikipedia	5231 Kb 5379 orfs MAP	1		50.5	NC_004431 AE014075	Univ of Wisconsin	NIAID	IMG NCBI Univ of Wisconsin TIGR	PNAS 99, 17020-4 2002-12-09
Gc00046	*Escherichia coli O157:H7* Sakai (EHEC)	B	PROTEOBACTERIA-GAMMA Taxonomy Entrez Wikipedia	5598 Kb 5253 orfs MAP	1	1	50.5	NC_002695	Japanese Consortium RIKEN	JSPS	IMG NCBI GIRC GIB TIGR	DNA Research 8, 11-22 2001-02-27
Gc00043	*Escherichia coli O157:H7* EDL933, EHEC	B	PROTEOBACTERIA-GAMMA Taxonomy Entrez Wikipedia	5528 Kb 5324 orfs MAP	1	2	50.4	NC_002655 AE005174	Univ of Wisconsin	NHGRI NIAID	IMG NCBI GIB TIGR	Nature 409,529-533 2001-01-25
Gc00008	*Escherichia coli* K12- MG1655	B	PROTEOBACTERIA-GAMMA Taxonomy Entrez EcoCyc Wikipedia	4639 Kb 4243 orfs MAP	1		50.8	NC_000913 U00096	Univ of Wisconsin	NHGRI	IMG NCBI SWISS EcoCyc Colibri IBM-Annotation EcoGene TIGR	Science 277,1453-1474 1997-09-05

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases



Genome databases - Genome pages at EBI

Archaea, Bacteria, Eukaryota, Organelle, Phage, Plasmid, Viroid, Virus

Allows proteomic searches (search for proteins in the genome)

Also possible to access data of ongoing projects

Files in EMBL format

164	Escherichia coli 536	4,938,920	CP000247	CP000247	Proteome
species Escherichia coli APEC O1					
165	Escherichia coli APEC O1	5,082,025	CP000468	CP000468	Proteome
species Escherichia coli CFT073 (Description)					
166	Escherichia coli CFT073	5,231,428	AE014075	AE014075	Proteome
species Escherichia coli K12 (Description)					
167	Escherichia coli K12 MG1655	4,639,675	U00096	U00096	Proteome
species Escherichia coli O157:H7 EDL933 (Description)					
168	Escherichia coli O157:H7 EDL933	5,528,445	AE005174	AE005174	Proteome
species Escherichia coli O157:H7 str. Sakai					
169	Escherichia coli O157:H7 str. Sakai	5,498,450	BA000007	BA000007	Proteome
species Escherichia coli UTI89					
170	Escherichia coli UTI89	5,065,741	CP000243	CP000243	Proteome
species Escherichia coli W3110					
171	Escherichia coli W3110	4,646,332	AP009048	AP009048	Proteome

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases

Genome databases - Genome pages at NCBI

Archaea, Bacteria, Eukaryota, Phage, Plasmid, Viroid, Virus

Also possible to access data of ongoing projects

Files in GenBank format

Escherichia coli 536		NC_008253	4938920 bp Jul 24 2006
Escherichia coli APEC O1		NC_008563	5082025 bp Nov 8 2006
Escherichia coli CFT073		NC_004431	5231428 bp Dec 9 2002
Escherichia coli K12		NC_000913	4639675 bp Jan 23 1997
Escherichia coli O157:H7	plasmid pO157	NC_007414	92077 bp Nov 3 1998
Escherichia coli O157:H7 EDL933		NC_002655	5528445 bp Sep 27 2001
Escherichia coli O157:H7 str. Sakai		NC_002695	5498450 bp Dec 4 2003
Escherichia coli O157:H7 str. Sakai	plasmid pOSAK1	NC_002127	3306 bp Jul 23 1998
Escherichia coli O157:H7 str. Sakai	plasmid pO157	NC_002128	92721 bp Jul 23 1998
Escherichia coli UTI89	plasmid pUTI89	NC_007941	114230 bp Apr 7 2006
Escherichia coli UTI89		NC_007946	5065741 bp Apr 7 2006
Escherichia coli W3110		AC_000091	4646332 bp Mar 1 2006

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

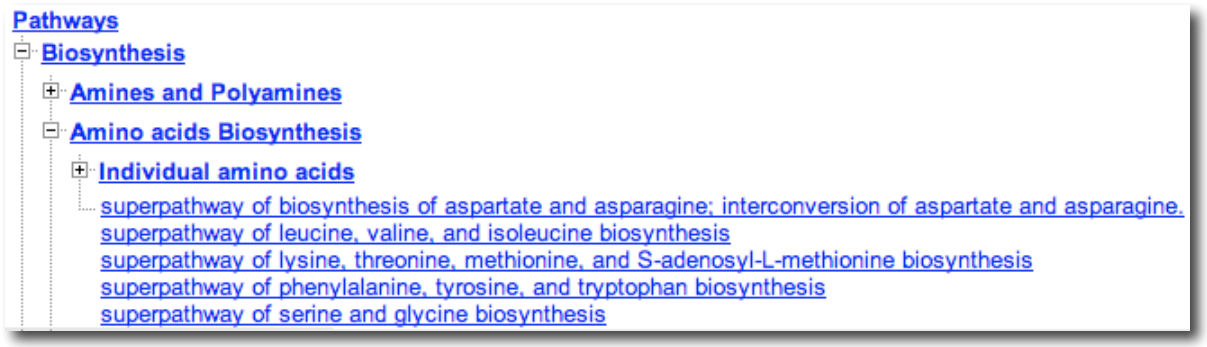
Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

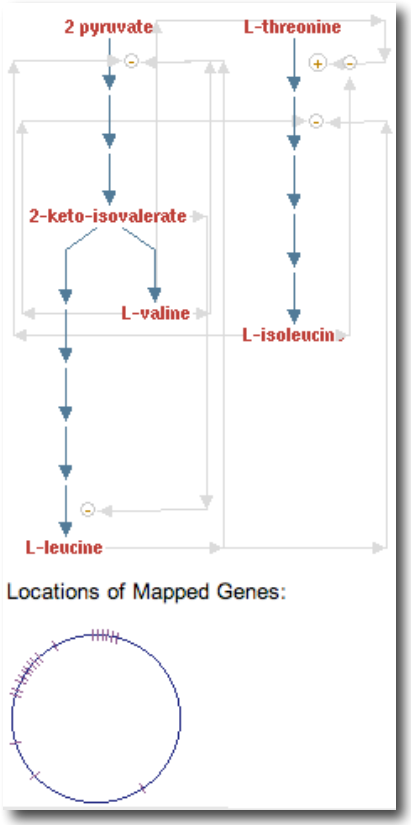
Genome and organism specific databases

Organism databases - *E. Coli* (http://ecocyc.org/)

Literature-based curation of the entire genome, and of transcriptional regulation, transporters, and metabolic pathways

Browse pathways:





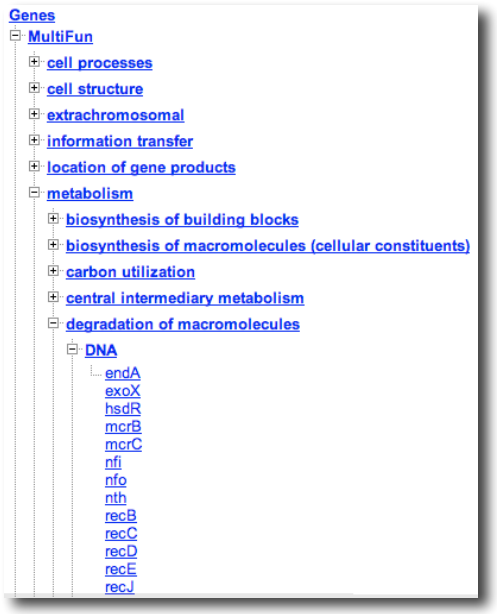
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases

Organism databases - *E. Coli* (http://ecocyc.org/)

Browse genes:





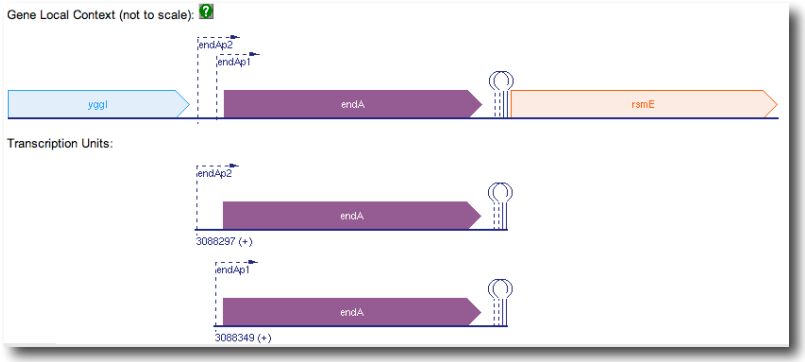
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

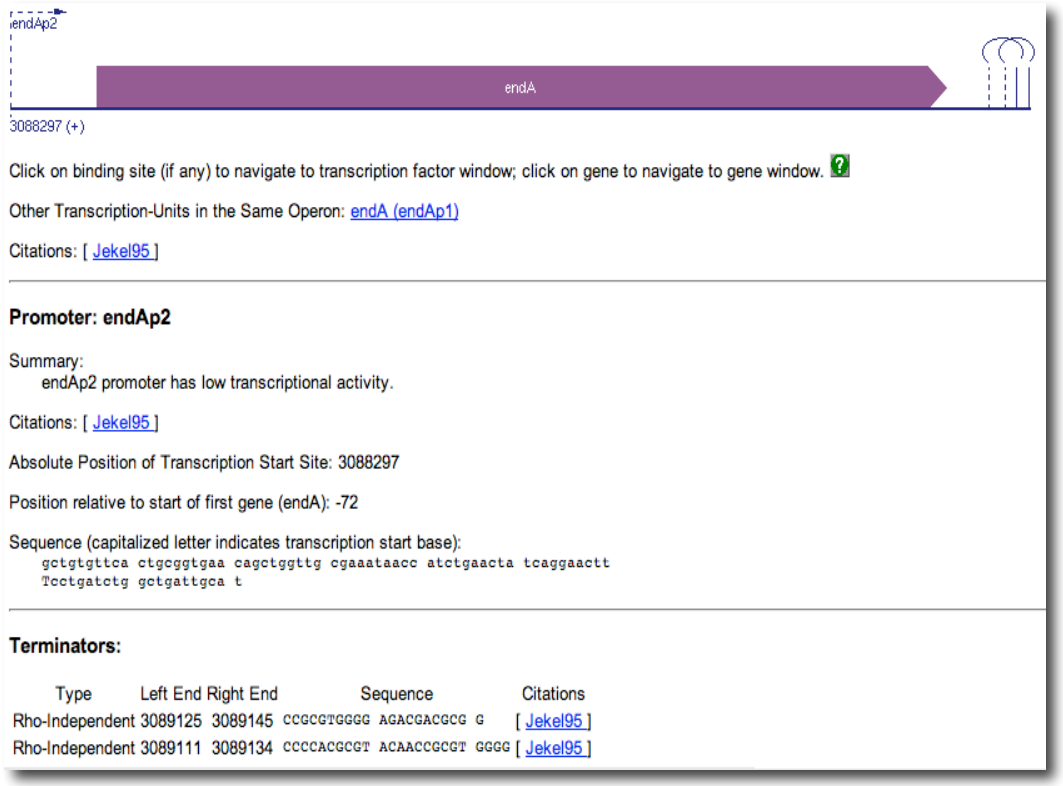
Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases

Organism databases - *E. Coli* (http://ecocyc.org/)

Genome browser:





Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

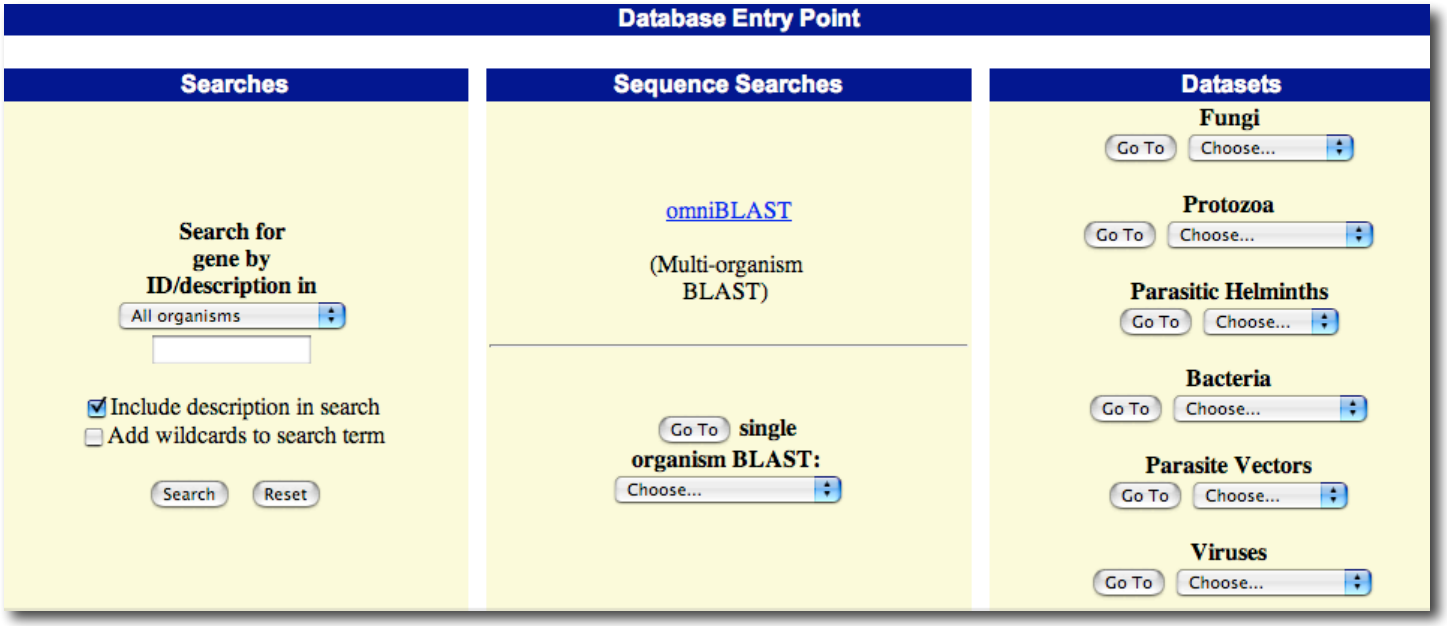
Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases

Organism databases - GeneDB

Browse several genomes sequenced at the Sanger Institute (PSU)

Integrated data from functional genomics and proteomics projects



Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases

Organism databases - GeneDB

E.g. Catalase from the plant pathogen *Erwinia carotovora*

General Information Add to Basket View Basket

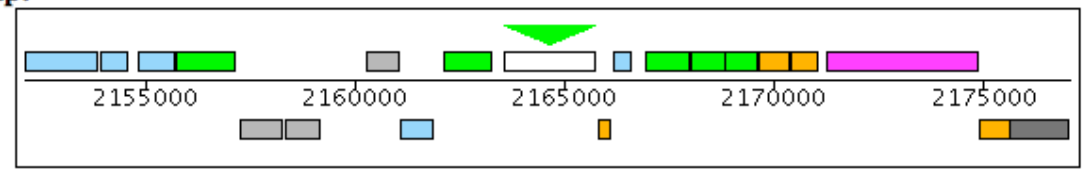
Name	katB
Systematic Name	ECA1909
Product	catalase-peroxidase
Type	CDS
Sequence	DNA and Protein

Location

Chromosome	1
Location	2163569..2165743 Length: 2175 bp

Region download and display (in Artemis)

Context Map:



narQ narP ECA1902 ECA1903 ECA1904 ECA1905 ECA1906 ECA1907 ECA1908 >katB< ECA1910
ECA1911 ECA1912 ECA1913 ECA1914 ECA1915 ECA1916 ECA1917 ECA1918 ECA1919

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Genome and organism specific databases

Organism databases - GeneDB

Primary Annotation

Similar to Legionella pneumophila catalase-peroxidase KatB SWALL:Q9ZGM4 (EMBL:AF078110) (721 aa) fasta scores: E(): 0, 78.28% id in 723 aa, and to Synechococcus sp. catalase-peroxidase SWALL:Q9R6S9 (EMBL:AF197161) (720 aa) fasta scores: E(): 0, 78.75% id in 720 aa

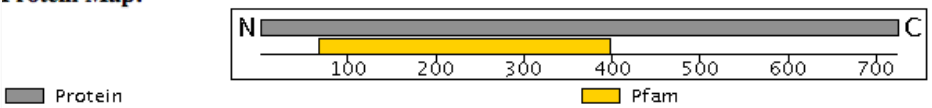
Predicted Peptide Properties

Mass	80.4 kDa	**Amino acids**	724
Isoelectric point	pH 5.1	**Charge**	-15.0

Signal Peptide Not found

Transmembrane Domains 0 found

Protein Map:



Domain Information

DB	Accs	Description	Note
Pfam	PF00141	Peroxidase	HMMPfam hit to PF00141, DE Peroxidase, score 4.1e-87

Database Similarities

DB	Acc	Organism	Description	Value	id	Overlap	Algorithm
UniProt	Q9ZGM4	*Legionella pneumophila*	catalase-peroxidase KatB	0	78.28%	723 aa	fasta
UniProt	Q9R6S9	*Synechococcus sp.*	catalase-peroxidase	0	78.75%	720 aa	fasta

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

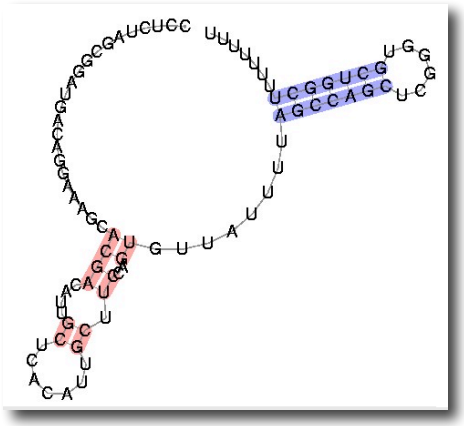
Other important databases

RNA database - Rfam

Large collection of multiple sequence alignments and covariance models covering many common non-coding RNA families

Contain information about RNA families and species distribution

RyhB RNA is a 90 nucleotide non-coding RNA that down-regulates a set of iron-storage and iron-using proteins when iron is limiting; it is itself negatively regulated by the ferric uptake repressor protein, Fur (Ferric uptake regulator). This ncRNA gene was recently identified in a screen and called SraI and was found to be expressed only in stationary phase.[2] RyhB RNA levels are inversely correlated with mRNA levels for the sdhCDAB operon, encoding succinate dehydrogenase, as well as five other genes previously shown to be positively regulated by Fur by an unknown mechanism. These include two other genes encoding enzymes in the tricarboxylic acid cycle, acnA and fumA, two ferritin genes, ftnA and bfr, and a gene for superoxide dismutase, sodB.[3] A number of other genes have been predicted computationally and verified as targets by microarray analysis: napF, sodA, cysE, yciS, acpS, nagZ and dadA.[1] RyhB is bound by the Hfq protein, that increases its interaction with its target messages.



```
BX950851.1/4651458-4651375 Erw.car.  ACGAUCCGGCUUCCCUGC..GAAGGCACGACAUUGCUCACAUUGCUUCCAGUAUUAUUU..AGCCAGC...UCGGGUGC
BX950851.1/2247623-2247706 Erw.car.  CCUCGCAGACACUGAAG...GAGAGCACGACAUUGCUCACAUUGCUUCCAGUAUUAUUUU.AGCCAGC...CUACGUGC
X76000.1/241-323           Pec.car.  UCUCGCAGGCACUGAAG...GAGAGCACGACAUUGCUCACAUUGCUUCCAGUAUUAUUUU.AGCCAGC...CGACGUGC
BX571859.1/98976-98893     Pho.lum.  GCCUUCGGGGU..ACAGAU.GAAAGCACGACAUUGCUCACAUUGCUUCCAGUGUUUACUUUAGCCAGC...CAAG.UGC
BX571867.1/241280-241198   Pho.lum.  GCUUUU.GUUAUCUCAGG..AAAAGCACGACAUUGCUCACAUUGCUUCCAGUGUUUUUU..AGCCAGC...UCGGGUGC
AE008863.1/16690-16610     Sal.typ.  CCCCUA.CGGAGAACCU...GAAAGCACGACAUUGCUCACAUUGCUUCCAGUAUUAUUUU.GCCCAGC...UUU..UGC
AE008755.1/9039-8957       Sal.typ.  UUUAUAACCCA..CCUG...UUCGGUACGACAUUGCUCACAUUGCUUCCAGUAUUAUUU..GCCCGCCUCACCGAGU.G
AL627281.1/157193-157273   Sal.ent.  CCCCUA.CGGAUAACCU...GAAAGCAUGACAUUGCUCACAUUGCUUCCAGUUUUAUUUU.GCCCAGC...UUU..UGC
AY183121.1/4-87            Yer.int.  CUUCUGGCUUUAGACAG...GAAAGCACGACAUUGCUCACAUUGCUUCCAGUGUUUUUUA.AGCCAGC...UCGGGUGC
AJ414160.1/4851-4766       Yer.pes.  AGUUUAGGUGUUCACAUUAUGAAAGCACGACAUUGCUCACAUUGCUUCCAGUGUUUACUU.AGCCAGC...CGGG.UGC
AY008264.2/524-607         Yer.ent.  UUUCAGGCUUUUGACAG...AAAAGCACGACAUUGCUCACAUUGCUUCCAGUGUUUUUUA.AGCCAGC...UCGGGUGC
AE017135.1/61734-61653     Yer.pes.  UCUUUGGCUUUGGUCAG...AAAAGCACGACAUUGCUCACAUUGCUUCCAGUAUU..UUA.AGCCAGC...UCGGGUGC
AE004101.1/4810-4886       Vib.cho.  ACUUGAGCAGGUUCUUU...UU.GACACGACAUUGCUCACAUUGCUUCCAGUGUAAUUUU.UAGCUUU...UUGG.UAA
```

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Other important databases

Pathway database - KEGG

Reference knowledge database linking genomes to biological systems

Several databases:
- GENES (Genomic information)
- LIGAND (Chemical information)
- PATHWAY and BRITE (Systems information)

Search KEGG for catalase Go Clear

Database	Release Info	Search & Compute	DBGET Search
KEGG PATHWAY	New maps Update status	Search objects in pathways Color objects in pathways	PATHWAY
KEGG BRITE	New hierarchies Update status	Map relations to hierarchies KEGG Orthology (KO)	BRITE
KEGG GENES	New organisms Update status	SSDB search BLAST search FASTA search EGassembler for ESTs KAAS automatic annotation GENIES network prediction	ORTHOLOGY GENES DGENES / EGENES GENOME VGENES / OGENES VGENOME
KEGG LIGAND	Update status	SIMCOMP compound search KCaM glycan search e-zyme reaction prediction PathComp computation	LIGAND COMPOUND DRUG GLYCAN REACTION RPAIR ENZYME

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

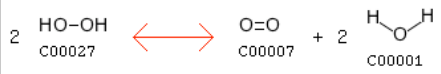
Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Other important databases

Pathway database - KEGG

GENES

Entry	EC 1.11.1.6 Enzyme
Name	catalase; equilase; caperase; optidase; catalase-peroxidase; CAT
Class	Oxidoreductases; Acting on a peroxide as acceptor; Peroxidases BRITE hierarchy
Sysname	hydrogen-peroxide:hydrogen-peroxide oxidoreductase
Reaction(IUBMB)	2 H2O2 = O2 + 2 H2O [RN:R00009]
Reaction(KEGG)	R00009; (other) R00602 R02670 Show all
Substrate	H2O2 [CPD:C00027]
Product	O2 [CPD:C00007]; H2O [CPD:C00001]
Cofactor	Heme [CPD:C00032]; Manganese [CPD:C00034]
Comment	A hemoprotein. This enzyme can also act as a peroxidase (EC 1.11.1.7 peroxidase) for which several organic substances, especially ethanol, can act as a hydrogen donor. A manganese protein containing MnIII in the resting state, which also belongs here, is often called pseudocatalase. Enzymes from some microorganisms, such as Penicillium simplicissimum, which exhibit both catalase and peroxidase activity, have sometimes been referred to as catalase-peroxidase.
Pathway	PATH: map00380 Tryptophan metabolism PATH: map00680 Methane metabolism PATH: map05030 Amyotrophic lateral sclerosis (ALS)
Orthology	KO: K00429 catalase KO: K03781 catalase KO: K03782 catalase/peroxidase

Entry	R00009 Reaction
Name	hydrogen-peroxide:hydrogen-peroxide oxidoreductase
Definition	2 H2O2 <=> Oxygen + 2 H2O
Equation	2 C00027 <=> C00007 + 2 C00001
	
RPair	RP: A02902 C00007_C00027 main
Enzyme	1.11.1.6
LinkDB	All DBs

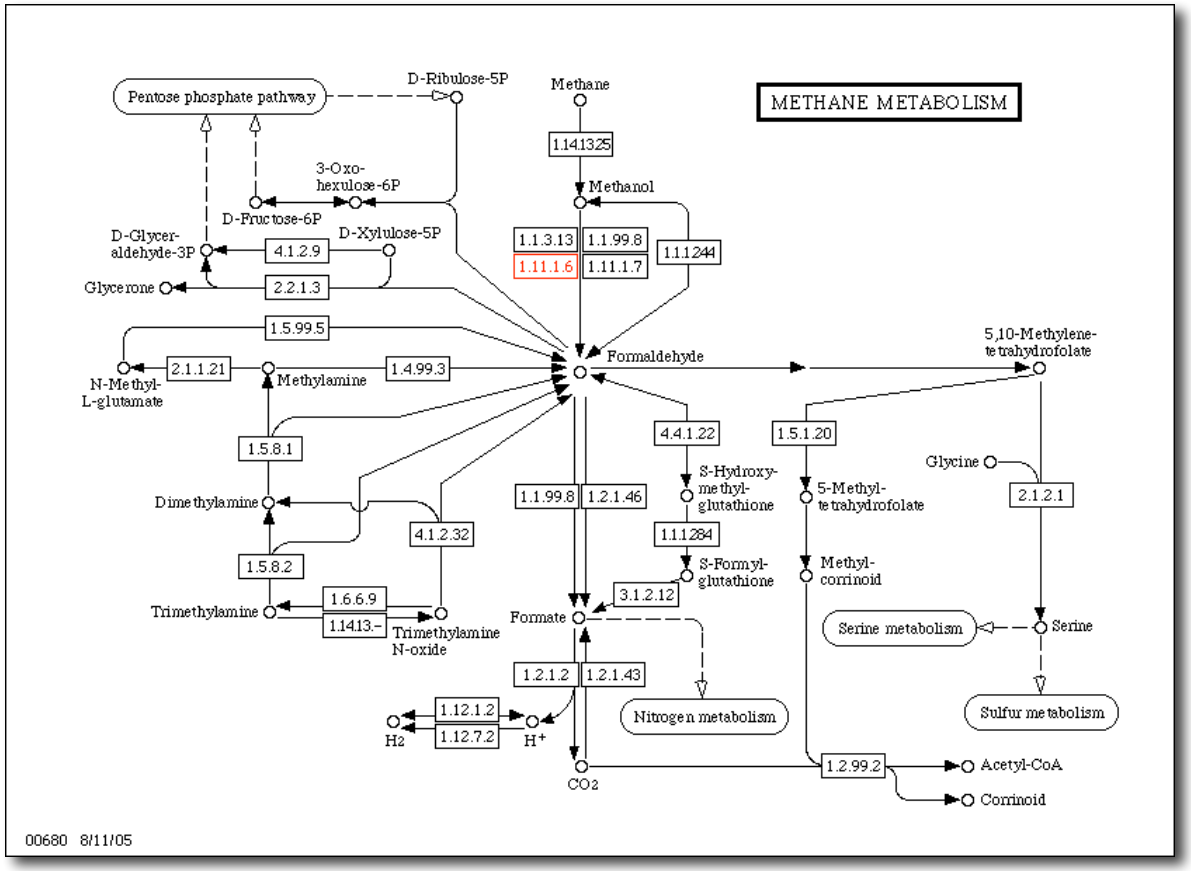
Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Other important databases

Pathway database - KEGG

PATHWAY



Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

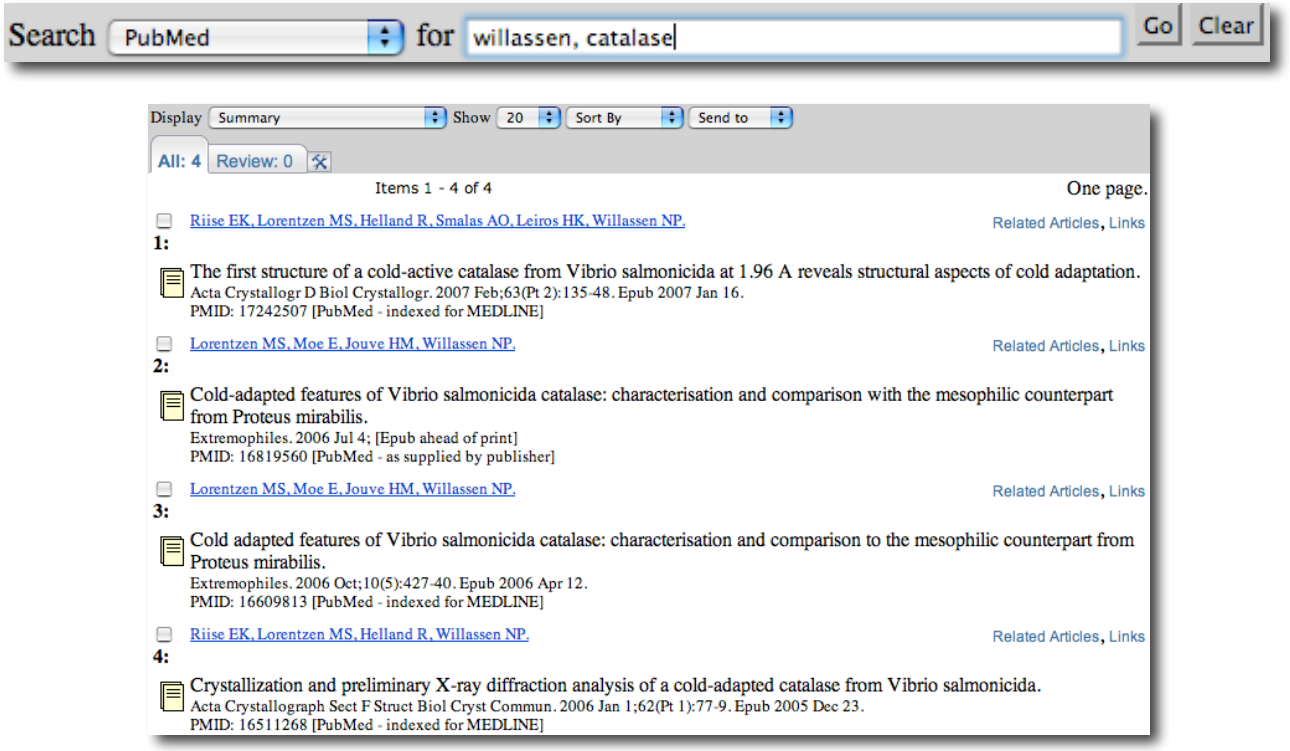
Other important databases

PubMed
A service of the National Library of Medicine and the National Institutes of Health
www.pubmed.gov

Literature database - PubMed at NCBI

Search database by words (e.g. author, year, or part of the title)

Contain abstracts of publications and direct links to the journals



Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway

Other important databases

PubMed
A service of the National Library of Medicine and the National Institutes of Health
www.pubmed.gov

Literature database - PubMed at NCBI

Contain abstracts of publications and direct links to the journals

Access full text if the paper is free or if you subscribe the particular journal

1: Extremophiles. 2006 Oct;10(5):427-40. Epub 2006 Apr 12.

Cold adapted features of Vibrio salmonicida catalase: characterisation and comparison to the mesophilic counterpart from Proteus mirabilis.

Lorentzen MS, Moe E, Jouve HM, Willassen NP.

Department of Molecular Biotechnology, Faculty of Medicine, Institute of Medical Biology, University of Tromsø, 9037 Tromsø, Norway.

The gene encoding catalase from the psychrophilic marine bacterium Vibrio salmonicida LFI1238 was identified, cloned and expressed in the catalase-deficient Escherichia coli UM2. Recombinant catalase from V. salmonicida (VSC) was purified to apparent homogeneity as a tetramer with a molecular mass of 235 kDa. VSC contained 67% heme b and 25% protoporphyrin IX. VSC was able to bind NADPH, react with cyanide and form compounds I and II as other monofunctional small subunit heme catalases. Amino acid sequence alignment of VSC and catalase from the mesophilic Proteus mirabilis (PMC) revealed 71% identity. As for cold adapted enzymes in general, VSC possessed a lower temperature optimum and higher catalytic efficiency (k (cat)/K (m)) compared to PMC. VSC have higher affinity for hydrogen peroxide (apparent K (m)) at all temperatures. For VSC the turnover rate (k (cat)) is slightly lower while the catalytic efficiency is slightly higher compared to PMC over the temperature range measured, except at 4 degrees C. Moreover, the catalytic efficiency of VSC and PMC is almost temperature independent, except at 4 degrees C where PMC has a twofold lower efficiency compared to VSC. This may indicate that VSC has evolved to maintain a high efficiency at low temperatures.

Related Links

- Cold-adapted features of Vibrio salmonicida catalase: characterisation and co [Extremophiles. 2006]
- The first structure of a cold-active catalase from Vibrio salmonicid [Acta Crystallogr D Biol Crystallogr. 2007]
- Crystallization and preliminary X-ray diffraction a [Acta Crystallograph Sect F Struct Biol Cryst Commun. 2006]
- Comparative studies of endonuclease I from cold-adapted Vibrio salmonicida and mesoph [FEBS J. 2007]
- High-resolution structure and biochemical properties of a recombinant Proteus mirabilis cat [Proteins. 2003]
- See all Related Articles...

Erik Hjerde
E-mail:erikh@fagmed.uit.no
Telephone: + 47 77 64 62 63
Fax: + 47 77 64 53 50

Protein Group
Department of Molecular Biotechnology
Faculty of Medicine
University of Troms, Norway