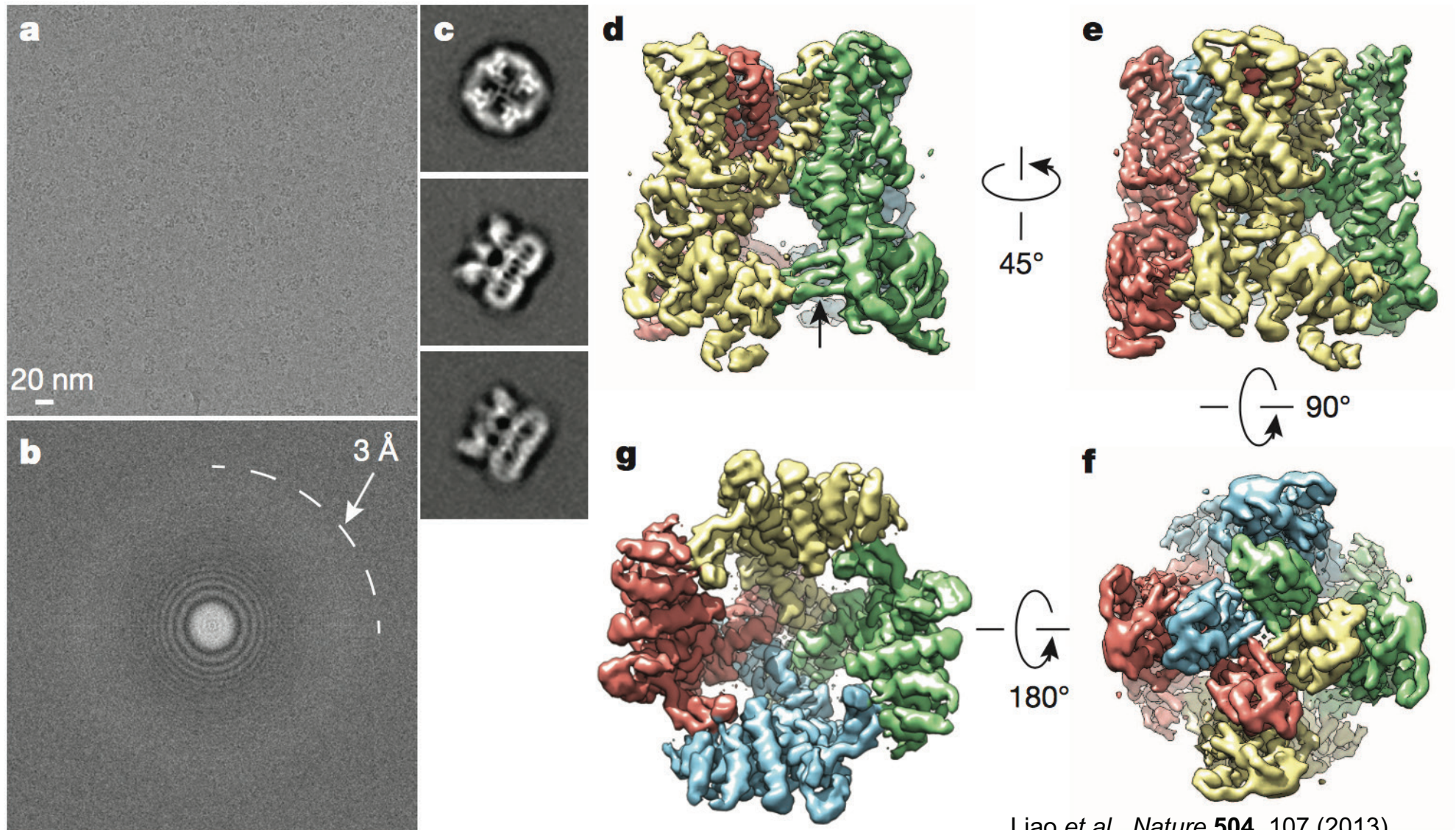


# Cryo-electron microscopy



Liao *et al.*, *Nature* **504**, 107 (2013)

- TRPV1 receptor (receptor for capsaicin – making chili “hot”)
- 3.4 Å resolution – breaking side-chain resolution barrier (PDB: 3J5P)

# Protein Structure Database

Jon K. Lærdahl,  
Structural Bioinformatics

Protein Data Bank (PDB) [www.rcsb.org](http://www.rcsb.org):

*The home of all experimental proteins structures*

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

An Information Portal to 124430 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands

Go

Advanced Search | Browse by Annotations

WORLDWIDE PDB PROTEIN DATA BANK

EMDataBank

Structural Biology Knowledgebase

Worldwide Protein Data Bank Foundation

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

**A Structural View of Biology**

This resource is powered by the Protein Data Bank archive—information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

**Discovering Biology Through Crystallography**

DISCOVERING BIOLOGY THROUGH CRYSTALLOGRAPHY COLORING BOOK

**November Molecule of the Month**

Aminopeptidase 1 and Autophagy

**Latest Entries** As of Tuesday Nov. 15

5M05 PDB Entry

X-ray crystal structure of myosin

**Features & Highlights**

**View Validation in 3D**  
Visualizing structure quality metrics in three dimensions • 10/11

**Explore Ligand Interactions in 3D**  
Analyze small molecule interactions with NGL • 10/11

**New Images for Transmembrane Proteins**  
Access multiple high resolution images that highlight orientation in membranes • 10/11

**News** Publications

**PDB and RCSB PDB: Did You Know?**  
Did you know PDB data are downloaded ~1.5 million times/day? Or that users can visualize the sites of genetic mutations with RCSB PDB tools? Download the State of the RCSB PDB for an overview of recent statistics and activities. • 11/15

Crossword Puzzle: Sequence Events • 11/08

Soon 135,000  
structures  
Not all are unique

Some few 1000  
unique protein folds

126,551,501,141  
bases in  
135,440,924  
sequence records in  
the traditional  
GenBank divisions  
as of April 2011


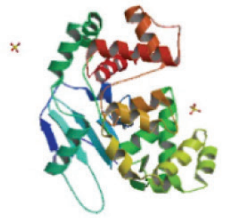
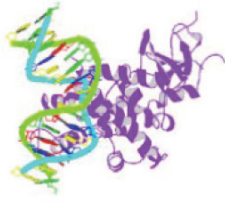
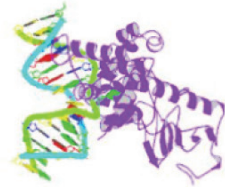
PDB identifiers are  
on the form 1LYZ,  
2B6C, 1T06 (and  
does not “mean”  
anything)



# Protein Structure Database

Jon K. Lærdahl,  
Structural Bioinformatics

Search for  
“OGG1”

 <a href="#">3D View</a>	<b>1LWY</b> <b>hOgg1 Borohydride-Trapped Intermediate without 8-oxoguanine</b> <a href="#">Fromme, J.C., Bruner, S.D., Yang, W., Karplus, M., Verdine, G.L.</a> (2003) Nat Struct Biol <b>10</b> 204-211 <b>Released:</b> 2/25/2003 <b>Method:</b> X-ray Diffraction <b>Resolution:</b> 2.01 Å <b>Residue Count:</b> 354 <b>Macromolecule:</b> 8-OXOGUANINE DNA GLYCOSYLASE (protein) <b>Unique Ligands:</b> PED	<a href="#">Download File</a> <a href="#">View File</a> <input checked="" type="checkbox"/>
 <a href="#">3D View</a>	<b>1KO9</b> <b>Native Structure of the Human 8-oxoguanine DNA Glycosylase hOGG1</b> <a href="#">Bjoras, M., Seeberg, E., Luna, L., Pearl, L.H., Barrett, T.E.</a> (2002) J Mol Biol <b>317</b> 171-177 <b>Released:</b> 1/9/2002 <b>Method:</b> X-ray Diffraction <b>Resolution:</b> 2.15 Å <b>Residue Count:</b> 345 <b>Macromolecule:</b> 8-oxoguanine DNA glycosylase (protein) <b>Unique Ligands:</b> SO4	<a href="#">Download File</a> <a href="#">View File</a> <input checked="" type="checkbox"/>
 <a href="#">3D View</a>	<b>1FN7</b> <b>COUPLING OF DAMAGE RECOGNITION AND CATALYSIS BY A HUMAN BASE-EXCISION DNA REPAIR PROTEIN</b> <a href="#">Norman, D.P., Bruner, S.D., Verdine, G.L.</a> (2001) J Am Chem Soc <b>123</b> 359-360 <b>Released:</b> 4/21/2001 <b>Method:</b> X-ray Diffraction <b>Resolution:</b> 2.6 Å <b>Residue Count:</b> 347 <b>Macromolecule:</b> 8-OXOGUANINE DNA GLYCOSYLASE 1 (protein) <b>Unique Ligands:</b> 3DR, CA	<a href="#">Download File</a> <a href="#">View File</a> <input checked="" type="checkbox"/>
 <a href="#">3D View</a>	<b>1EBM</b> <b>CRYSTAL STRUCTURE OF THE HUMAN 8-OXOGUANINE GLYCOSYLASE (HOGG1) BOUND TO A SUBSTRATE OLIGONUCLEOTIDE</b> <a href="#">Bruner, S.D., Norman, D.P., Verdine, G.L.</a> (2000) Nature <b>403</b> 859-866 <b>Released:</b> 3/20/2000 <b>Method:</b> X-ray Diffraction <b>Resolution:</b> 2.1 Å <b>Residue Count:</b> 347 <b>Macromolecule:</b> 8-OXOGUANINE DNA GLYCOSYLASE (protein) <b>Unique Ligands:</b> 8OG, CA	<a href="#">Download File</a> <a href="#">View File</a> <input checked="" type="checkbox"/>

# Protein Structure Database

Jon K. Lærdahl,  
Structural Bioinformatics

First hit for  
"OGG1"

Structure Summary 3D View Annotations Sequence Sequence Similarity Structure Similarity Experiment Literature

PDB id

1KO9

Native Structure of the Human 8-oxoguanine DNA Glycosylase hOGG1

DOI: [10.2210/pdb1ko9/pdb](https://doi.org/10.2210/pdb1ko9/pdb)

Classification: [HYDROLASE](#)

Deposited: 2001-12-20 Released: 2002-01-09

Deposition author(s): [Bjoras, M.](#), [Seeberg, E.](#), [Luna, L.](#), [Pearl, L.H.](#), [Barrett, T.E.](#)

Organism: [Homo sapiens](#)

Expression System: Escherichia coli

Structural Biology Knowledgebase: 1KO9 (1 model >24 annotations) [SBKB.org](#)

Display Files Download Files

PDB file (data file)

View structure in e.g. JSmol

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 2.15 Å

R-Value Free: 0.252

R-Value Work: 0.206

wwPDB Validation

Full Report

Metric	Percentile Ranks	Value
Clashscore		12
Ramachandran outliers		0
Sidechain outliers		2.7%

Worse Percentile relative to all X-ray structures Better Percentile relative to X-ray structures of similar resolution

Literature

Download Primary Citation

Reciprocal "flipping" underlies substrate recognition and catalytic activation by the human 8-oxoguanine DNA glycosylase.

[Bjoras, M.](#), [Seeberg, E.](#), [Luna, L.](#), [Pearl, L.H.](#), [Barrett, T.E.](#)

(2002) J.Mol.Biol. 317: 171-177

PubMed: 11902 Search on PubMed

DOI: [10.1006/jmb.2002.5400](https://doi.org/10.1006/jmb.2002.5400)

PubMed Abstract

Both 8-oxo-guanine and formamidopyrimidines are major products of oxidative DNA damage that can result in the fixation of transversion mutations following replication if left unrepaired. These lesions are targeted by the N-DNA glycosylase hOgg1, which catalyses excision of the aberrant

View in 3D: JSmol or PV (in Browser)

Standalone Viewers

[Simple Viewer](#) [Protein Workshop](#)  
[Ligand Explorer](#) [Kiosk Viewer](#)

Protein Symmetry: Asymmetric ([View in 3D](#))

Protein Stoichiometry: Monomer

Biological assembly 1 assigned by authors

Macromolecular Content

Unique protein chains: 1

Publication

Resolution

# PDB entry – an example in PDB format

- Standard since early 1970s
- FORTRAN compatible format
- Some limitations
  - Number of atoms
  - Number of chains
  - Length of fields
- Not good for parsing by computers

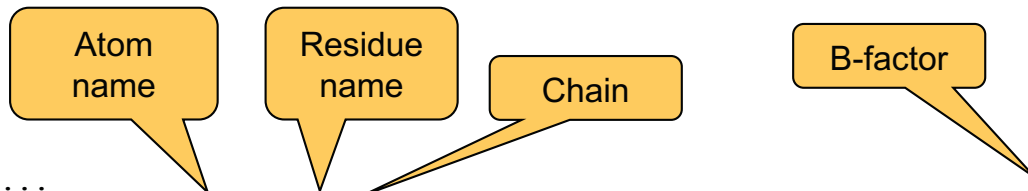
```

HEADER      LYASE/DNA                                24-JAN-00    1EBM
TITLE       CRYSTAL STRUCTURE OF THE HUMAN 8-OXOGUANINE GLYCOSYLASE
TITLE       2 (HOGG1) BOUND TO A SUBSTRATE OLIGONUCLEOTIDE
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;
COMPND      3 CHAIN: A;
COMPND      4 FRAGMENT: CORE FRAGMENT (RESIDUES 12 TO 325);
COMPND      5 SYNONYM: AP LYASE;
COMPND      6 ENGINEERED: YES;
COMPND      7 MUTATION: YES;
COMPND      8 MOL_ID: 2;
COMPND      9 MOLECULE: DNA (5'-D(*GP*CP*GP*TP*CP*CP*AP*(OXO)
COMPND     10 GP*GP*TP*CP*TP*AP*CP*C)-3');
COMPND     11 CHAIN: C;
COMPND     12 ENGINEERED: YES;
COMPND     13 MOL_ID: 3;
COMPND     14 MOLECULE: DNA (5'-
COMPND     15 D(*GP*GP*TP*AP*GP*AP*CP*CP*TP*GP*GP*AP*CP*GP*C)-3');
COMPND     16 CHAIN: D;
COMPND     17 ENGINEERED: YES
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE      3 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE      4 EXPRESSION_SYSTEM_COMMON: BACTERIA;
SOURCE      5 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE      6 EXPRESSION_SYSTEM_PLASMID: PET30A-HOGG1;
SOURCE      7 MOL_ID: 2;
SOURCE      8 SYNTHETIC: YES;
SOURCE      9 MOL_ID: 3;
SOURCE     10 SYNTHETIC: YES
KEYWDS      DNA REPAIR, DNA GLYCOSYLASE, PROTEIN/DNA
EXPDTA      X-RAY DIFFRACTION
AUTHOR      S.D.BRUNER,D.P.NORMAN,G.L.VERDINE
REVDAT      1    20-MAR-00 1EBM    0
JRNL         AUTH    S.D.BRUNER,D.P.NORMAN,G.L.VERDINE
JRNL         TITL    STRUCTURAL BASIS FOR RECOGNITION AND REPAIR OF THE
JRNL         TITL 2  ENDOGENOUS MUTAGEN 8-OXOGUANINE IN DNA
JRNL         REF     NATURE                                V. 403    859 2000
JRNL         REFN    ASTM NATUAS    UK ISSN 0028-0836
REMARK      1
REMARK      2 RESOLUTION. 2.10 ANGSTROMS.
REMARK      3
.....

```

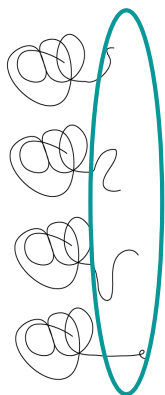
# PDB entry – an example in PDB format

Jon K. Lærdahl,  
Structural Bioinformatics



Amino acid field

The B-factor (temperature factor) is an indicator of thermal motion. Actually a mixture of real thermal motion and structural disorder (multiple conformations)



Cofactor field

....												
ATOM	1	N	GLY	A	9	29.382	-12.935	38.434	1.00	39.96		N
ATOM	2	CA	GLY	A	9	28.983	-13.096	36.994	1.00	40.83		C
ATOM	3	C	GLY	A	9	27.548	-12.643	36.792	1.00	41.51		C
ATOM	4	O	GLY	A	9	27.265	-11.724	36.007	1.00	41.29		O
ATOM	5	N	SER	A	10	26.631	-13.287	37.505	1.00	41.40		N
ATOM	6	CA	SER	A	10	25.222	-12.936	37.418	1.00	41.42		C
ATOM	7	C	SER	A	10	24.900	-11.903	38.494	1.00	39.54		C
ATOM	8	O	SER	A	10	23.732	-11.620	38.763	1.00	40.12		O
ATOM	9	CB	SER	A	10	24.357	-14.176	37.639	1.00	43.12		C
ATOM	10	OG	SER	A	10	24.599	-14.728	38.920	1.00	43.93		O
ATOM	11	N	GLU	A	11	25.940	-11.343	39.102	1.00	37.35		N
ATOM	12	CA	GLU	A	11	25.764	-10.360	40.166	1.00	36.30		C
ATOM	13	C	GLU	A	11	26.373	-9.013	39.755	1.00	34.00		C
ATOM	14	O	GLU	A	11	27.302	-8.968	38.951	1.00	32.56		O
ATOM	15	CB	GLU	A	11	26.451	-10.849	41.454	1.00	38.36		C
ATOM	16	CG	GLU	A	11	26.387	-12.365	41.740	1.00	39.94		C
ATOM	17	CD	GLU	A	11	25.069	-12.823	42.343	1.00	41.33		C
ATOM	18	OE1	GLU	A	11	24.963	-14.021	42.693	1.00	40.98		O
ATOM	19	OE2	GLU	A	11	24.139	-11.999	42.468	1.00	41.16		O
ATOM	20	N	GLY	A	12	25.853	-7.925	40.320	1.00	31.94		N
ATOM	21	CA	GLY	A	12	26.368	-6.602	40.009	1.00	30.07		C
ATOM	22	C	GLY	A	12	25.925	-6.027	38.674	1.00	29.09		C
ATOM	23	O	GLY	A	12	25.174	-6.652	37.919	1.00	28.15		O
ATOM	24	N	HIS	A	13	26.392	-4.820	38.379	1.00	29.23		N
ATOM	25	CA	HIS	A	13	26.043	-4.159	37.124	1.00	29.36		C
ATOM	26	C	HIS	A	13	26.651	-4.913	35.941	1.00	30.04		C
ATOM	27	O	HIS	A	13	27.838	-5.247	35.948	1.00	30.64		O
ATOM	28	CB	HIS	A	13	26.545	-2.716	37.121	1.00	28.62		C
ATOM	29	CG	HIS	A	13	25.874	-1.831	38.127	1.00	27.87		C
ATOM	30	ND1	HIS	A	13	26.285	-1.746	39.441	1.00	26.37		N
....												
HETATM	3056	O	HOH		5	23.168	15.174	34.624	1.00	18.07		O
HETATM	3057	O	HOH		6	21.609	14.592	31.635	1.00	13.68		O
HETATM	3058	O	HOH		7	14.739	30.965	30.601	1.00	26.62		O
HETATM	3059	O	HOH		9	29.320	3.836	25.672	1.00	27.62		O
....												

Atom coordinates

# PDB entry – an example in mmCIF format

Newer data format and  
alternative to “PDB format”

- No limitations in number of atoms, chains, fields etc.
- Better suited for automatic parsing/processing

```
data_1EBM
#
_entry.id      1EBM
#
_audit_conform.dict_name      mmcif_pdbx.dic
_audit_conform.dict_version   1.044
_audit_conform.dict_location  http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.
_database_2.database_code
PDB  1EBM
NDB  PD0117
RCSB RCSB010437
#
_database_PDB_rev.num          1
_database_PDB_rev.date         2000-03-20
_database_PDB_rev.date_original 2000-01-24
_database_PDB_rev.status       ?
_database_PDB_rev.replaces      1EBM
_database_PDB_rev.mod_type      0
#
_pdbx_database_status.status_code REL
_pdbx_database_status.entry_id    1EBM
_pdbx_database_status.deposit_site RCSB
_pdbx_database_status.process_site RCSB
_pdbx_database_status.SG_entry    .
#
loop_
_audit_author.name
'Bruner, S.D.'
'Norman, D.P.'
'Verdine, G.L.'
#
_citation.id          primary
_citation.title        'Structural basis for recognition'
_citation.journal_abbrev Nature
_citation.journal_volume 403
_citation.page_first    859
_citation.page_last     866
```

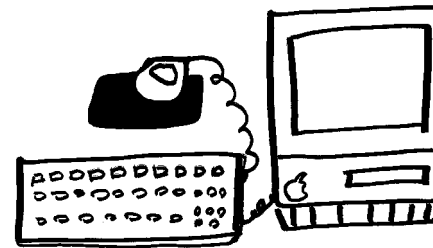
# Structural bioinformatics

Jon K. Lærdahl,  
Structural Bioinformatics

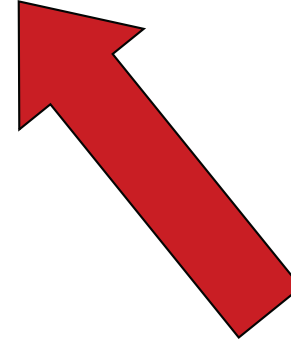
Experimental structure is hard to get

The 3D structure on a protein is determined by the amino acid sequence (primary structure)

There are many orders of magnitude more sequences available than there are structures



How do we get information about structure from sequence?





# Protein domains

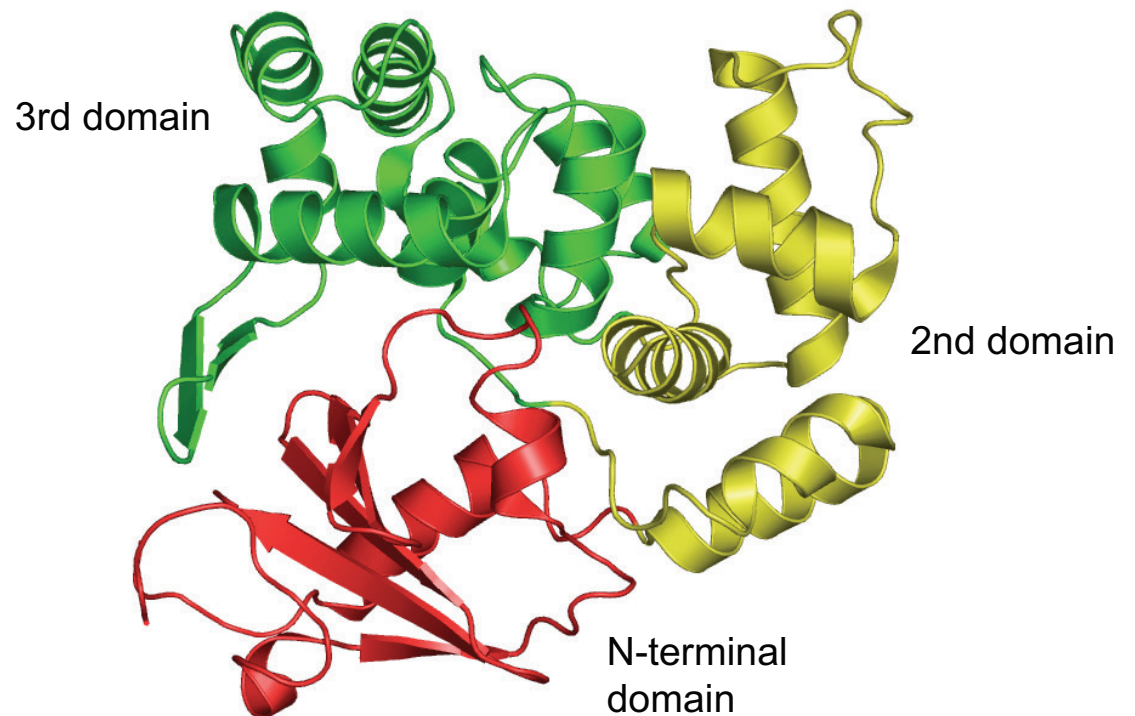
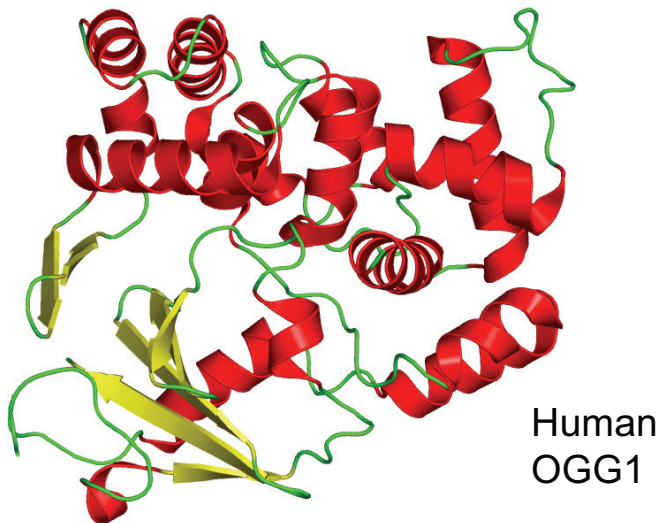
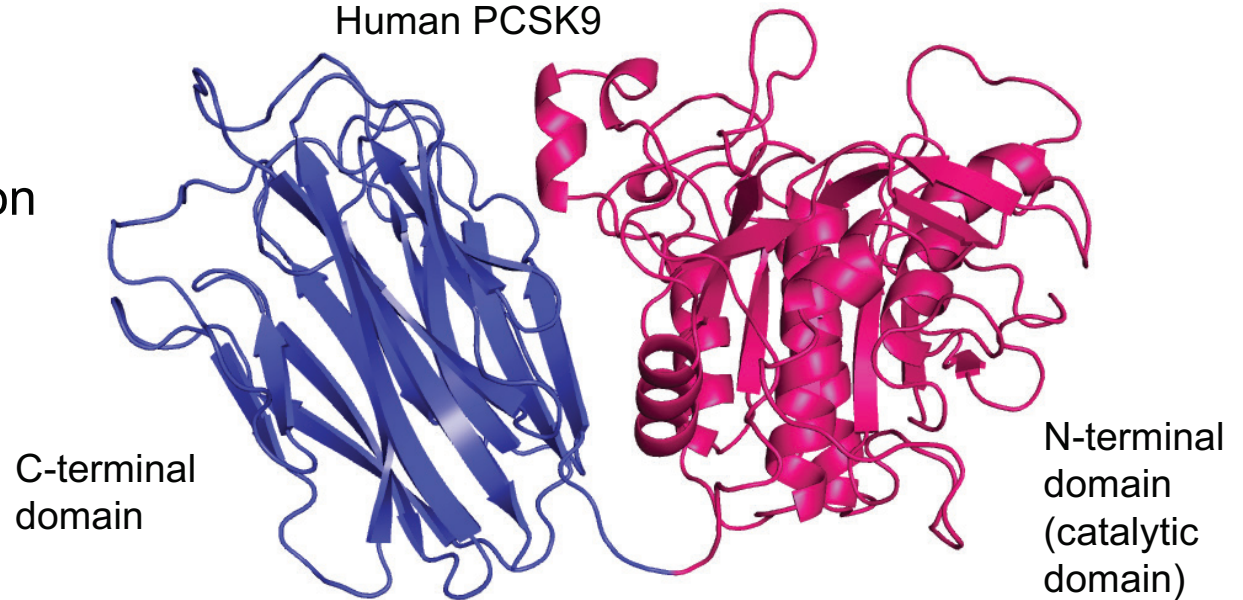
Jon K. Lærdahl,  
Structural Bioinformatics

**Domain:** Compact part of a protein that represents a structurally independent region

Domains are often separate functional units that may be studied separately

Domains fold independently?  
Not always...

Human PCSK9



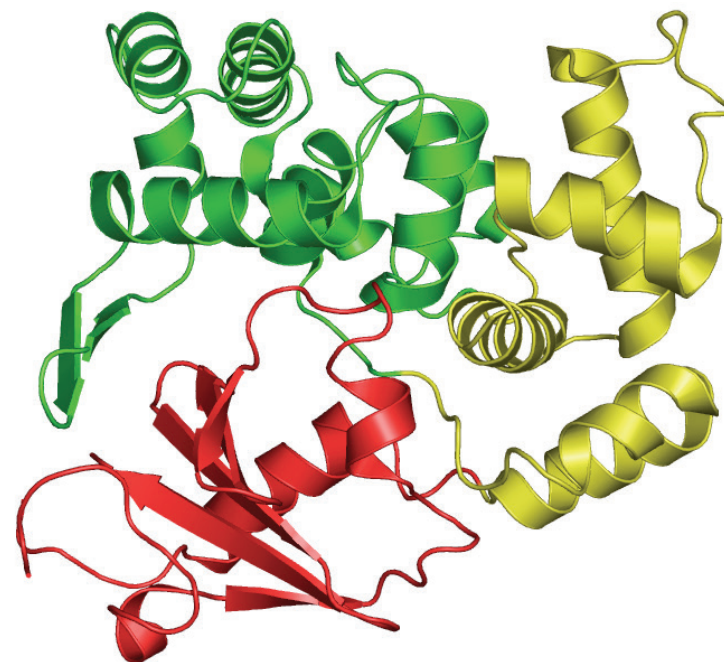
# Protein domains

Jon K. Lærdahl,  
Structural Bioinformatics

Dividing a protein structure into domains: no “right way to do it” or “correct algorithm”, *i.e. a lot of subjectivity involved*



Most people would agree there are two domains here



Three domains?  
One domain?  
Two?

SCOP vs.  
CATH?

Very often we model, compare, classify *domains* – not full-length proteins

# Protein domains

Jon K. Lærdahl,  
Structural Bioinformatics

Instead of working with full length proteins that may be

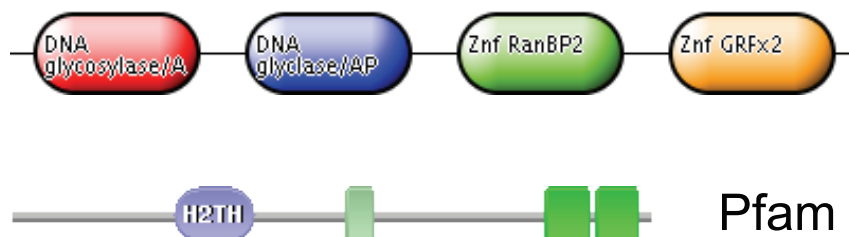
- very large
- contain one or many separate modules (*i.e.* domains)
- have both structured and unstructured parts

We often instead work with protein domains that are

- more compact
- can be studied separately
  - function
  - structure by X-ray crystallography/NMR
  - bioinformatics modeling
- may be viewed as the “spare parts” building up full-length proteins

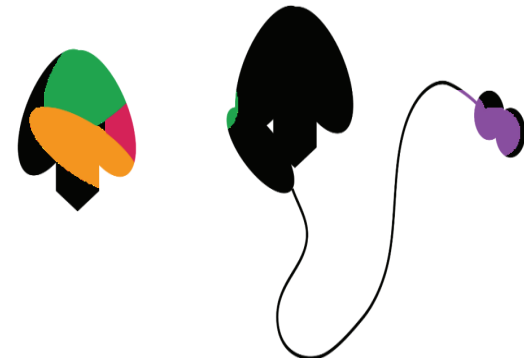
Many proteins are structured domains, “spare parts”, connected by short loops or long disordered regions

Far from trivial to detect boundaries between domains from sequence only:



InterPro

Pfam



# Protein domains

Jon K. Lærdahl,  
Structural Bioinformatics

Domains have a “signature sequence” that can be described as a HMM Logo

Domains can be “switched”. They can be viewed as “spare parts” that can be used to build new proteins through evolution

**Important to think in terms of domains!!**

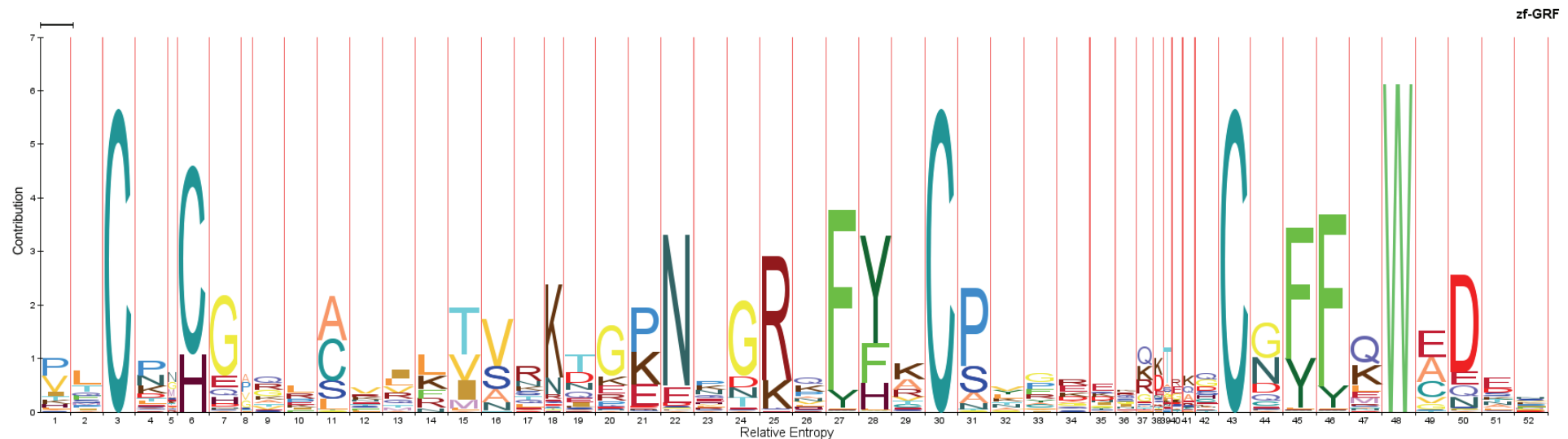
 GRF zinc finger domain

Human NEIL3 

Human APEX2 

Human Topoisomerase IIIα





Pfam HMM-logo for the GRF zinc finger domain



## Nature of the protein universe

PNAS **106**, 11079 (2009)

Michael Levitt<sup>1</sup>

Department of Structural Biology, Stanford University, Stanford, CA 94305-5126

Contributed by Michael Levitt, May 9, 2009 (sent for review April 20, 2009)

The protein universe is the set of all proteins of all organisms. Here, all currently known sequences are analyzed in terms of families that have single-domain or multidomain architectures and whether they have a known three-dimensional structure. Growth of new single-domain families is very slow: Almost all growth comes from new multidomain architectures that are combinations of domains characterized by  $\approx 15,000$  sequence profiles. Single-domain families are mostly shared by the major groups of organisms, whereas multidomain architectures are specific and account for species diversity. There are known structures for a quarter of the single-domain families, and  $>70\%$  of all sequences can be partially modeled thanks to their membership in these families.

featured in a recent report on the Protein Structure Initiative (7) that expressed concern that because the number of new families is expanding rapidly determining three-dimensional structures for a representative of each family may not be possible (8).

Here, we approach the problem differently. Instead of clustering entire protein sequences (6), we rely on the occurrence of protein sequence patterns termed “sequence profiles.” These patterns can be derived from a few members of the family and then used to add new members that match the same pattern.

An obvious way to cluster sequences into families is by pairwise comparison (4) of all sequences preceded by indexing (5) to eliminate close pairs. Such a combination led to massive clustering of millions of protein sequences from both known species and environmental samples by Yooseph et al. (6). Their remarkable conclusion was that the number of protein families as measured by the number of sequence clusters showed no sign of saturation. Indeed, the cluster count was increasing at the same rate as new sequences were being determined. This result

(6) Yooseph D, *et al.* (2007) The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. PLoS Biol **5**:e16.

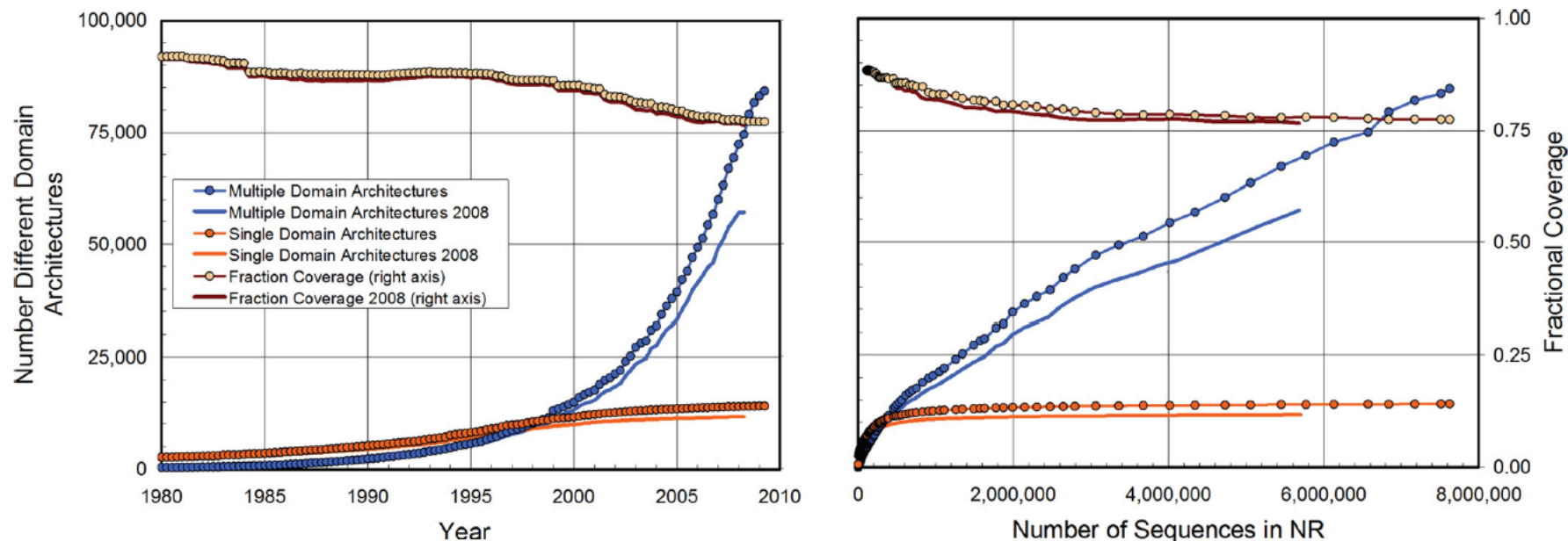
[www.pnas.org/cgi/doi/10.1073/pnas.0905029106](http://www.pnas.org/cgi/doi/10.1073/pnas.0905029106)



# Protein domains

Jon K. Lærdahl,  
Structural Bioinformatics

PNAS **106**, 11079 (2009)



**Fig. 1.** As the NR database grows, the number of different multidomain architecture (MDA) families found by CDART is increasing rapidly with year (*Left*) or added sequence (*Right*). In contrast, the number of single-domain architecture (SDA) families is increasing much more slowly. Because the number of sequences is growing exponentially, fractional sequence coverage (number of sequences in a SDA or MDA family divided by the total number of NR sequences) has dropped slightly from 0.88 to 0.76; more than three-quarters of current sequences still contain a domain recognized by a known sequence profile. Merged CDART sequence profiles are used here. Corresponding results with unmerged CDART sequence profiles are given in [Fig. S1](#). The solid curves marked “2008” were made with a release of CDART from February 9, 2008, which contained fewer sequence profiles (24,083 compared with 27,036). This gave rise to smaller numbers of SDA and MDA families and lower coverage. During this time, the number of sequences in the NR database increased by 2 million.

There are known structures for a quarter of the single-domain families, and >70% of all sequences can be partially modeled thanks to their membership in these families.

End