# Protein Common Interface Database (ProtCID)

## Overview

ProtCID contains comprehensive, PDB-wide structural information on the interactions of proteins and individual protein domains with other molecules, including four types of interactions: chain interfaces, Pfam domain interfaces, Pfam-peptide interfaces and Pfam-ligand/nucleic acids interactions. A common interaction here indicates chain-chain or Pfam domain-domain interfaces that occur in different crystal forms or Pfam-peptide or Pfam-ligand interactions that occur in multiple homologous proteins.

Its main goal is to identify and cluster homodimeric and heterodimeric interfaces observed in multiple crystal forms of homologous proteins, and interactions of peptide and ligands in homologous proteins. Such interfaces and interactions, especially of non-identical proteins or protein complexes, have been associated with biologically relevant interactions. For more details about the algorithm and benchmarking, please refer to our paper "Statistical Analysis of Interface Similarity in Crystals of Homologous Proteins." and the ProtCID web site (http://dunbrack2.fccc.edu/ProtCiD).

## Repo Contents

### Source code

C# code for 10 ProtCID dynamic link libraries (dll folders), including

1. AuxFuncLib

The library contains helper functions: zip or unzip files, parse or write PDB files, run linux programs remotely, format SQL query string, a lot more.

1. BuCompLib

The library is used to process biological assemblies from PDB and PISA, including chain interfaces, domain interfaces, chain-peptide interfaces, chain-ligand interactions, and comparison of PDB and PISA biological assemblies of PDB entries. This library is used to generate a firebird database named bucomp.fdb.

1. BuQueryLib

The library is to query biological assemblies from bucomp.fdb database, and format biological assemblies in four formats: ABC format (e.g. PDB: 3FYU, A3B2C, the chain with maximum copies is always named “A”), entity format (e.g. (1.1)(2.2)(3.3) in the entity ID order in the PDB entry file), asymmetric chain format (e.g. (A)(B, D)(C, E, F) in the entity ID order in the PDB entry file), author chain format (e.g. (A)(B, D)(C, E, F) in the entity ID order in the PDB entry file).

1. CrystalInterfaceLib

This is one of main library. It contains all source code to generate interfaces from crystal structures, including build 3x3x3 unit cells from PDB asymmetric unit files, compute interfaces by K-DOPs algorithm, calculate similarity scores and surface area values.

1. DataCollectorLib

The library is to generate all sequence and structural alignments at chain and Pfam domain levels, including FatCat structural alignments (<http://fatcat.sanfordburnham.org/>), HH alignments from HHsuite (<https://github.com/soedinglab/hh-suite>), and sequence alignments from PsiBlast (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/).

1. DBLib

The library contains all operations on ProtCID databases, including database connection, table creation, and query/update/delete/insert data.

1. InterfaceClusterLib

This is one of main library. It contains all source code to cluster interfaces, including classify interfaces to homologous groups at chain level and domain level (Pfam), cluster interfaces, compile all coordinate files of clusters, and generate all PyMol scripts.

1. ProgressLib

The progress library is to provide a progress bar and window to show the progress of data processing.

1. ProtCidSettingsLib

This library is to set and get all directory settings and parameter settings.

1. ProtCidWebDataLib

This library is to generate meta data for ProtCID web site to speed up web queries.

1. Main program

The main window form is for all settings, all functions to build ProtCID databases, and progress bar and progress window. However, it is not feasible just click menu items to rebuild ProtCID databases on the entire PDB database and many other data sources used to build ProtCID. ProtCID contains hundreds GBs databases and millions of interface files, cluster files and text files. That is why we provide a web site http://dunbrack2.fccc.edu/protcid, so users can query on our database.

### Web Site

<http://dunbrack2.fccc.edu/protcid>

ProtCID web site contains two services of Windows Communication Foundation (WCF). One is to query on the ProtCID database, and the other one is to assign Pfams to input sequences. The details about ProtCID and how to use it, please refer the HELP pages on ProtCID web site.

### Demo

This folder contains a Windows console program to generate chain interfaces on a list of PDB entries, cluster interfaces, output interface files in PDB format, output result text files (e.g. similarity Q scores), compile coordinates of each cluster including PyMol scripts to visualize each cluster.

This demo program uses ProtCID libraries, and can be installed by the Windows installer.

## System Requirements

All source code of ProtCID including the demo program is written in C# in Visual studio 2013. It requires Windows operating system and .NET Framework 4.5.

## Installation Guide

It is very easy to install the demo program. Just download protcid\_demo\_setup.msi or setup.exe (<https://github.com/DunbrackLab/ProtCID>/demo), double click the installer, follow the steps, but to change the installation directory where the program can read and write to it.

## Instructions for Use

#### Synopsis

Protcid\_demo –infile ls-pdb.txt –datadir datadir [options]

ProtCid\_demo generates interfaces for each PDB entry in “ls-pdb.txt”, clusters interfaces and stores all files in “datadir”.

-infile a text file containing a list of PDBs, one PDB per line

-datadir the path where result data are to be saved

#### Options

-alnfile a text file containing a multiple sequence alignment for the input PDBs. The program will use this file to map residues when calculating similarity Q score of two interfaces from different entries. If this file is not provided, the author residue numbers are used when calculating Q scores. The alignment can be clustal omega format, or a simple text file, one line for each PDB sequence or chain, gaps must be filled by ‘-’, like

clustal omega format

CLUSTAL O(1.2.4) multiple sequence alignment

1gwnC MGSSHHHHHHSSGLVPRGSHMDPNQNVKCKIVVVGDSQCGKTALLHVFAKDCFPENYVPT 60

5p21A -------------------------MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPT 35

1gwnC VFENYTASFEIDTQRIELSLWDTSGSPYYDNVRPLSYPDSDAVLICFDISRPETLDSVLK 120

5p21A IEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIH- 94

OR

1gwnC MGSSHHHHHHSSGLVPRGSHMDPNQNVKCKIVVVGDSQCGKTALLHVFAKDCFPENYVPTVFENYTASFEIDTQRIELSLWDTSGSPYYDNVRPLSYPDS DAVLICFDISRPETLDSVLK

5p21A -------------------------MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTG EGFLCVFAINNTKSFEDIH-

-groupname to create a folder so all results are saved into a specific folder with the name, and also named clusters. For instance, -groupname ras. All results are saved into a folder under “datadir”, and clusters are named by ras\_cluster ID.tar.gz, e.g. ras\_1.tar.gz for the first cluster of user group “ras”.

Examples to run ProtCid\_demo in Windows command line after change the directory of ProtCID\_demo.exe:

1. ProtCID\_demo –infile demo\_data\ls-pdb\_ST1A1.txt –datadir demo\_data

This will generate and cluster interfaces from PDB entries in ls-pdb\_ST1A1.txt, save all files and coordindates to demo\_data folder. The PDB entries contain same protein ST1A1\_HUMAN. When calculating Q scores, author residue numbers are used.

1. ProtCID\_demo –infile demo\_data\ls-pdb\_ST1A1.txt –datadir demo\_data –groupname **sulf**

This will generate and cluster interfaces from PDB entries in ls-pdb\_ST1A1.txt, same all files and coordinates to demo\_data\**sulf** folder and named cluster coordinates files in **sulf**\_cluster ID.tar.gz (e.g. **sulf**\_1.tar.gz).

1. ProtCID\_demo –infile demo\_data\ls-pdb\_RAS.txt –datadir demo\_data –alnfile demo\_data\ RasMonomers\_clustalO.aln –groupname ras

This will generate and cluster interfaces from PDB entries in ls-pdb\_RAS.txt, calculate similarity Q scores by updating all residue numbers to alignment positions (1:N of the alignment). The sequence alignment can be generated by Clustal omega or other sequence or structure alignment programs like PsiBlast or FATCAT.

## Results

Output of program include

1. PDB xml files downloaded from rcsb web site, and converted into ProtCID specific xml files. <folder: xml>
2. Text files

EntryInfo.txt, EntityChainInfo.txt, EntityUnpDbRef.txt (parsed from PDB xml files)

CrystInterfaces.txt (definition of interfaces computed from crystals)

SameEntryInterfaceCompInfo.txt, DiffEntryInterfaceCompInfo.txt (similarity Q scores between interfaces of each entry, between interfaces of two entries)

InterfaceClusters.txt (Clusters of interfaces)

1. Interface files (interface coordinate files in PDB format)
2. Cluster coordinates files (coordinates of each cluster including Pymol scripts to visualize a cluster )