The usefulness of integrated databases: a case study of COLUMBA

[1]Stefan Günther, [1]Kristian Rother, [2]Silke Trissl, [1]Elke Michalsky,
[1]Cornelius Froemmel, [2]Ulf Leser

[1]Institute for Biochemistry, Charite-Campus Mitte, Monbijoustr. 2, 10098 Berlin
[2]Institute for Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin

Corresponding author: stefan.guenther@charite.de

Protein Structure Annotation

www.columba-db.de

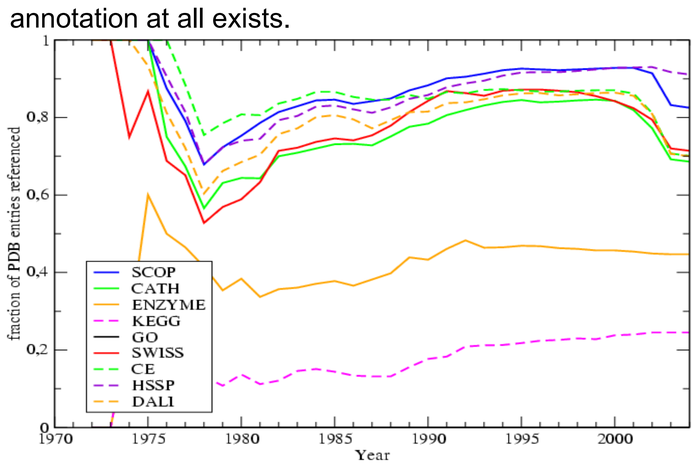
Database Content

The Columba database of protein structure annotation currently integrates the PDB and annotation from twelve second-party databases (Rother et al., 2004). Using Columba, users can quickly create subsets of the PDB given conditions on various sequence and structure features. We investigated to what extent PDB entries are annotated by second-party information based on existing cross-references between PDB and nine other databases.

Database	#entries	description
PDB	24932	protein structures
SCOP	52877	folding classification
CATH	48391	folding classification
SWISS-Prot	138779	protein sequences
ENZYME	4242	enzyme descriptions
KEGG	120	metabolic pathways
GOA	877096	gene ontology annotation
DALI	17468	automatic fold/family classification
CE	17522	automatic fold/family classification
HSSP	22712	automatic fold/family classification

Annotation is self-consistent but incomplete

Protein structures published before 1997 are in general well-annotated. Roughly two thirds of the PDB entries are linked to both a sequence database and one or more fold classifications. The fraction of PDB entries covered is quite high through most of the last decade. But after 1996, references to SWISS-Prot become less abundant. Approximately 77% of the PDB fall into that time range. The same applies for the folding classifications after 2000. For 2010 proteins published after 2000 (8% of the PDB), no annotation at all exists.



Fraction of the PDB that is covered by second-party annotation for a specific year. The fraction drops observably over the last years where human interaction is required, while it stays constant for automatically created annotation (KEGG, ENZYME and HSSP).

Methods

The availability of cross-references should be considered in the creation of representative sets, as a structure is more useful, the more secondary annotation is available. Choosing a set of structures based on properties that are only available in second-party annotations introduces a strong bias into the selection. This bias suppresses peptides, non-Proteins, unfolded and exotic structures, but also suppresses recent structures, where "recent" stretches over a period of approximately four years.

The lack of annotations especially for recent structures, which are often those of best quality, is certainly worrisome. Since the data production pace increases, it is likely that the gap will become larger rather than shrink. This gap can only be closed by increasing the amount of resources invested into manual inspection and annotation of protein structures, unless methods for automated structure annotation are improving significantly.

Refereces

Rother, K., Mueller, H., Trissl, S., Koch, I., Steinke, T., Preissner, R., Froemmel, C., and Leser, U. COLUMBA: Multidimensional Data Integration of Protein Annotations. *In Data Integration in the Life Sciences, (ed. Erhard Rahm.).* pp. 156 – 171. Springer, Leipzig, Germany. (2004)

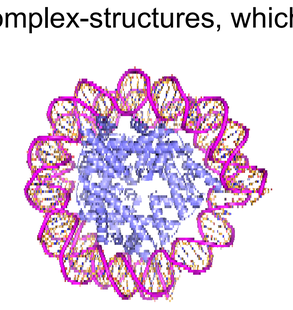
Hoppe, A., Frömmel, C. NeedleHaystack: A program for the rapid recognition of local structures in large sets of atomic coordinates. *J. Appl. Cryst.* **36**: 1090 – 1097. (2003)

Analysis of local similarity of DNA binding sites

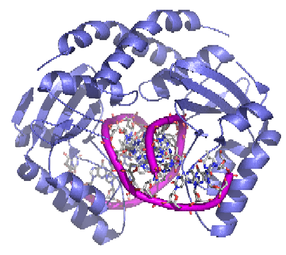
Introduction

Interactions between proteins and DNA play an essential role in gene expression, structuring, DNA-replication and repair. Up to the present the mechanism of DNA binding is not completely understood.

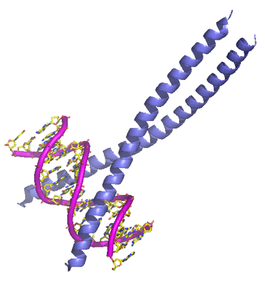
Sequence specific binding requires a certain local structure to allow different biochemical interactions between the atoms of the two biopolymers. To detect similar local binding sites we analyzed the contact regions of nearly all protein-DNA-complex-structures, which are present in the PDB.



Histone (1kx4)



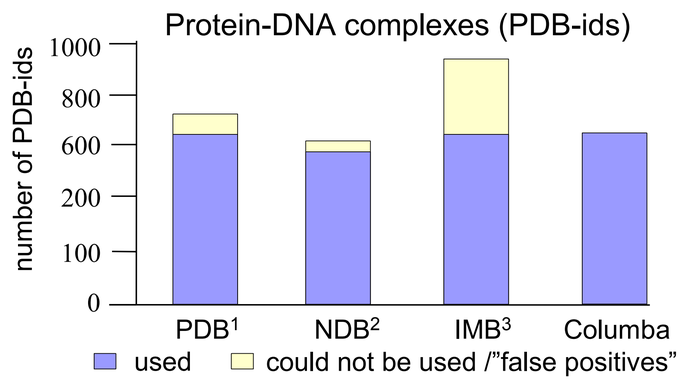
Endonuclease (1eyu)



Leucine zipper (1nwq)

Data Set

Different databases with collections of DNA-complex-IDs were cross-compared. In many cases the structures of the protein-DNA-complexes contain only short peptides or the DNA consists of only few nucleotides (“false positives”). The database Columba enabled us to apply detailed search criteria to sort out these PDB-ids. Additionally the database was used to extract the chains and divide them into clusters by sequence identity.



650 PDB-Structures of Protein-DNA-complexes

↓

1194 chains with DNA-contact

↓

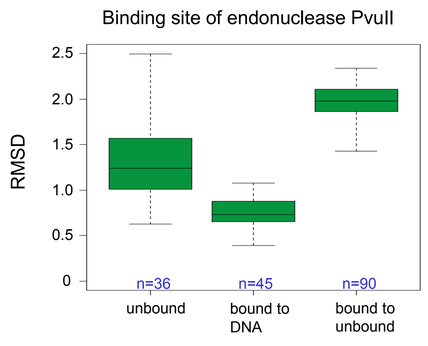
216 clusters, each with similar sequences (> 90% identity)

[1]Protein Data Bank – www.rcsb.org/pdb
[2]Nucleic Acid Database – ndbserver.rutgers.edu
[3]Jena Image Library of Biological Macromolecules www.imb-jena.de

Methods

Binding regions were defined as those protein atoms within a radius of 5 Å around the DNA. We compared the 3D structures of the binding regions with the Needle-Haystack search algorithm (Hoppe & Frömmel 2003). The algorithm picks out similar regions of two protein structures and calculate the root-mean-square-deviation (RMSD). The structural similarities of highly homologous binding sites with and without DNA as well as binding sites of different helix-turn-helix proteins were characterized.

Results



Within a cluster all binding sites were superimposed with each other.

We found that within a sequence cluster the DNA binding regions can be completely diverse as long as no DNA is bound to that region. On DNA binding, these binding sites are highly similar, but none resembles the unbound conformation.

Screening against the entire dataset with a helix-turn helix motif, we found many closely related proteins. However, the specificity of matches does not allow to reasonably identify distant relatives.

Charité Medical School and University Hospital, Humboldt-Universität zu Berlin
Humboldt-Universität zu Berlin
Konrad-Zuse-Zentrum für Informationstechnik Berlin
University of Applied Sciences, Berlin

This work was done within the context of the NGFN Nationales Genomforschungsnetz

and is supported by sponsored by bmb+f