

Chapter 2

Creating Datasets

Abstract Data is the key in biological knowledge discovery. The data used in discovery is specific and specialized to a specific issue in cell and molecular biology. This is generally achieved by creating datasets of specific nature. Here, we discuss the importance of biological datasets in information gleaned and describe procedures for specialized dataset creation. The creation of data subsets for human leukocyte antigen (HLA) peptide binding, HLA–peptide structures, HLA class I and class II grouping of structures with peptides, protein subunit interactions, homodimers, heterodimers, homodimer folding into categories, fusion proteins, intron-containing genes in eukaryotes and intronless genes in eukaryotes, is described in this chapter.

Keywords Data • Dataset • Subset • Source • Derived • Grouping • Class • Features • Analysis • Molecule specific • HLA • MHC • Peptide • Protein subunit • Interactions • Intron • Intronless • Folding

2.1 Datasets

Biological phenomena are specific, yet diverse in nature. Description of such phenomena using data requires specialized datasets derived from large databases such as GenBank/EMBL/DDBJ and PDB (Fig. 2.1). Databases mentioned here are the major resources for genetic information. However, the source of data is not limited to these databases. Creation of specialized datasets is important to understand issues in several disciplines (immunology, macromolecular structural biology, biochemistry and genetics) of molecular and cellular biology. Datasets of small scale (homogeneous data – similar data property) are created from huge databases (heterogeneous data – diverse data property) using keywords or annotations used to describe specific biological property or phenomenon. The dataset thus created is generally of redundant (unnecessary or duplicated) in nature.

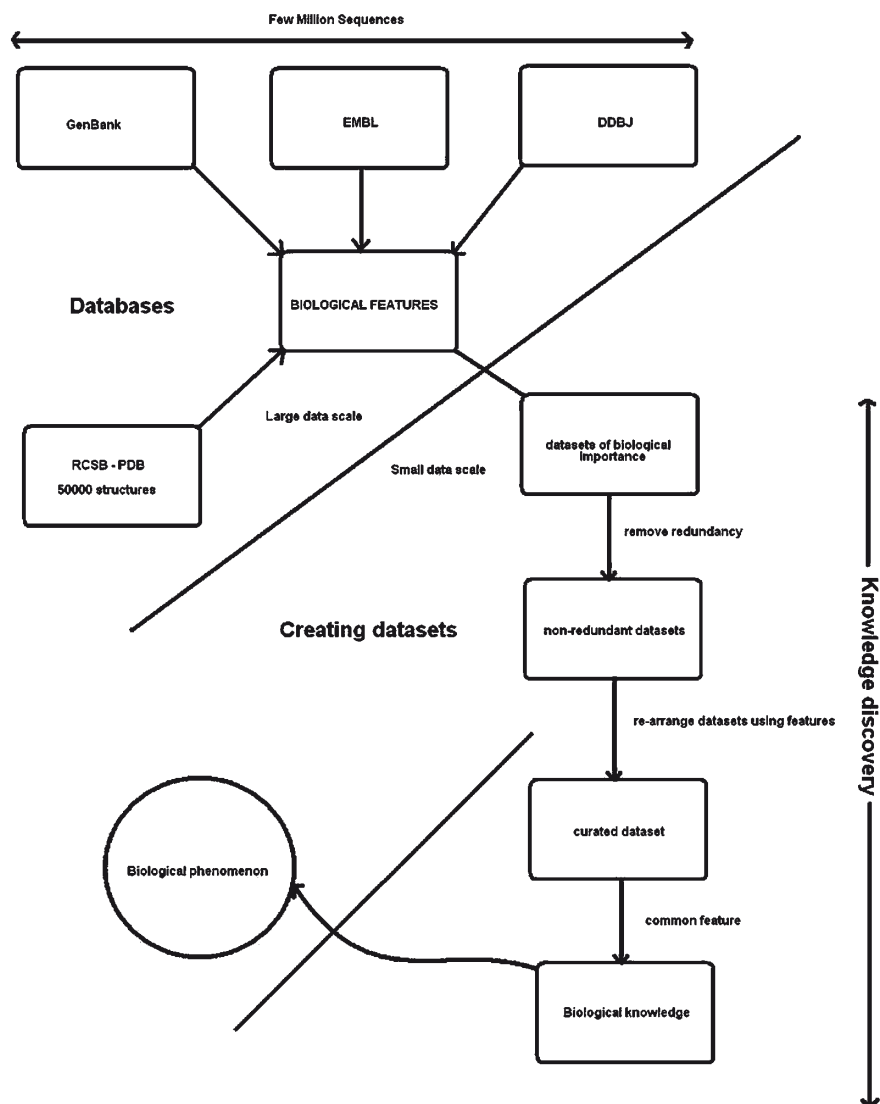


Fig. 2.1 Creating biological datasets for knowledge discovery. *PDB* protein databank, *DDBJ* DNA databank of Japan, *EMBL* European molecular biology laboratory, *RCSB* research collaboration for structural biology

Therefore, it is important to remove redundancy and to create a nonredundant dataset of interest. The nonredundant dataset is further curated (process of data refinement) to add value. This is done by data grouping or clustering. This procedure again clusters data of similar nature into specific subgroups. The subgroups

thus generated are helpful in identifying common features that are used to describe a specific biological phenomenon. Therefore, it is critical to create specialized datasets. In this chapter, the creation of a number of biologically important datasets is described.

2.2 HLA-Binding Peptide Dataset

Human leukocyte antigen (HLA) allele-specific peptide-binding data were collected from literature (Alexander et al. 1998; Den Haan et al. 1998; Kawashima et al. 1998; Chang et al. 1999; Nukaya et al. 1999). These binding data are strictly based on HLA molecules defined by the sequence-based nomenclature system described by IMGT/HLA (<http://www.ebi.ac.uk/imgt/hla/>). Any binding data defined by the serological nomenclature system is not included in our analysis. Binding values for these peptides were expressed in IC_{50} units, denoting the peptide concentration inhibiting the binding of the standard peptide by 50%. This provides an estimate of the ability of the peptides binding to specific HLA allele proteins. The collected peptides were grouped into different subsets based on the HLA alleles they bind (Table 2.1). Each of the subset was further clustered into three subgroups by their IC_{50} units: good binders – less than 100 nM; moderate binders 100–500 nM and poor binders or nonbinders more than 500 nM (Table 2.2). The grouping of data provides an insight to the type of short peptides that specifically bind or not bind to defined HLA allele antigens.

2.3 MHC–Peptide Structural Dataset

Major histocompatibility complex (MHC) plays an important role in T-cell-mediated immune response. The structural information on MHC–peptide complexes was retrieved from the Protein Databank (PDB) as flat files with the “.pdb” extension (www.rcsb.org/PDB/). The retrieved data amounts to 43 in number and this dataset was later used for further analysis (Table 2.3: class I dataset and Table 2.4: class II dataset). The currently available structural data on MHC–peptide complexes in PDB is partially curated and somewhat redundant in nature. A fully curated dataset for MHC–peptide complexes was created by visual inspection of the structural data which further led to the development of MHC–peptide interaction database (MPID – <http://surya.bic.nus.edu.sg/mpidt/>). This approach produces the most accurate and useful information for discovery given the current limitations of a purely automatic procedure. The rules applied to clean data, the types of rules governing MHC–peptide interactions are also discussed.

Table 2.1 HLA class I-specific peptides with IC₅₀ binding values

HLA allele	Peptide	IC ₅₀	References
A*0201	VHRDDLLEA	365	Kawashima et al. 1998
A*0202	KIFGSLAFL	9.0	Kawashima et al. 1998
A*0202	KVAELVHFL	29	Kawashima et al. 1998
A*0202	FLWGPRALV	43	Kawashima et al. 1998
A*0202	IMIGVLVGV	62	Kawashima et al. 1998
A*0202	YLSGANLNL	165	Kawashima et al. 1998
A*0202	LLTFWNPPV	297	Kawashima et al. 1998
A*0202	ILHNGAYSL	358	Kawashima et al. 1998
A*0202	VMAGVGSPYV	391	Kawashima et al. 1998
A*0202	CLTSTVQLV	457	Kawashima et al. 1998
A*0202	LMTFWNPPV	779	Kawashima et al. 1998
A*0202	LLTFWNPPPT	1,720	Kawashima et al. 1998
A*0202	ALCRWGLLL	>10,000	Kawashima et al. 1998
A*0203	YLSGANLNL	2.4	Kawashima et al. 1998
A*0203	CLTSTVQLV	6.7	Kawashima et al. 1998
A*0203	LMTFWNPPV	7.6	Kawashima et al. 1998
A*0203	KIFGSLAFL	9	Kawashima et al. 1998
A*0203	IMIGVLVGV	13	Kawashima et al. 1998
A*0203	VMAGVGSPYV	13	Kawashima et al. 1998
A*0203	FLWGPRALV	14	Kawashima et al. 1998
A*0203	KVAELVHFL	14	Kawashima et al. 1998
A*0203	LLTFWNPPV	26	Kawashima et al. 1998
A*0203	LLTFWNPPPT	67	Kawashima et al. 1998
A*0203	ILHNGAYSL	100	Kawashima et al. 1998
A*0203	ALCRWGLLL	278	Kawashima et al. 1998
A*0203	YLQLVFGIEV	345	Kawashima et al. 1998
A*0206	KIFGSLAFL	23	Kawashima et al. 1998
A*0206	LMTFWNPPV	33	Kawashima et al. 1998
A*0206	LLTFWNPPV	56	Kawashima et al. 1998
A*0206	IMIGVLVGV	106	Kawashima et al. 1998
A*0206	KVAELVHFL	168	Kawashima et al. 1998
A*0206	CLTSTVQLV	308	Kawashima et al. 1998
A*0206	FLWGPRALV	336	Kawashima et al. 1998
A*0206	YLQLVFGIEV	370	Kawashima et al. 1998
A*0206	ILHNGAYSL	567	Kawashima et al. 1998
A*0206	LLTFWNPPPT	755	Kawashima et al. 1998
A*0206	YLSGANLNL	804	Kawashima et al. 1998
A*0206	VMAGVGSPYV	3,700	Kawashima et al. 1998
A*0206	ALCRWGLLL	8,863	Kawashima et al. 1998
A*0301	RLGVRATRK	11.7	Chang et al. 1999
A*0301	QLFTFSPRR	14.7	Chang et al. 1999
A*0301	RMVYGGVEHR	15.3	Chang et al. 1999
A*0301	LIFCHSKKK	20.4	Chang et al. 1999
A*0301	GVAGALVAFK	28.2	Chang et al. 1999
A*0301	VAGALVAFK	45.8	Chang et al. 1999
A*0301	KTSESRQPR	68.8	Chang et al. 1999
A*0301	LGFGAYMSK	135.8	Chang et al. 1999
A*1101	LIFCHSKKK	1.6	Chang et al. 1999

(continued)

Table 2.1 (continued)

HLA allele	Peptide	IC ₅₀	References
A*1101	GVAGALVAFK	4.3	Chang et al. 1999
A*1101	VAGALVAFK	6.7	Chang et al. 1999
A*1101	LGFGAYMSK	20.7	Chang et al. 1999
A*1101	KTSESRQPR	93.8	Chang et al. 1999
A*1101	QLFTFSPRR	182.0	Chang et al. 1999
A*1101	RLGVRATRK	207.0	Chang et al. 1999
A*1101	RMVVGVEHR	300.0	Chang et al. 1999
A*3101	KTSESRQPR	66.7	Chang et al. 1999
A*3101	RMVVGVEHR	94.7	Chang et al. 1999
A*3101	RLGVRATRK	428.6	Chang et al. 1999
A*3101	QLFTFSPRR	620.7	Chang et al. 1999
A*3101	LIFCHSKKK	2535.6	Chang et al. 1999
A*3101	LGFGAYMSK	2950.8	Chang et al. 1999
A*3101	GVAGALVAFK	3272.7	Chang et al. 1999
A*3101	VAGALVAFK	3750.0	Chang et al. 1999
A*3301	KTSESRQPR	1812.5	Chang et al. 1999
A*3301	QLFTFSPRR	3766.2	Chang et al. 1999
A*3301	RMVVGVEHR	9666.7	Chang et al. 1999
A*3301	RLGVRATRK	10,000	Chang et al. 1999
A*3301	GVAGALVAFK	10,000	Chang et al. 1999
A*3301	LGFGAYMSK	10,000	Chang et al. 1999
A*3301	LIFCHSKKK	10,000	Chang et al. 1999
A*3301	VAGALVAFK	10,000	Chang et al. 1999
A*6801	QLFTFSPRR	2.6	Chang et al. 1999
A*6801	GVAGALVAFK	117.6	Chang et al. 1999
A*6801	KTSESRQPR	145.5	Chang et al. 1999
A*6801	LGFGAYMSK	222.2	Chang et al. 1999
A*6801	VAGALVAFK	258.1	Chang et al. 1999
A*6801	LIFCHSKKK	333.3	Chang et al. 1999
A*6801	RMVVGVEHR	1777.8	Chang et al. 1999
A*6801	RLGVRATRK	10,000	Chang et al. 1999
A*6802	KVAELVHFL	17	Kawashima et al. 1998
A*6802	FLWGPRALV	40	Kawashima et al. 1998
A*6802	IMIGVLVGV	89	Kawashima et al. 1998
A*6802	KIFGSLAFL	3,333	Kawashima et al. 1998
A*6802	LMTFWNPPV	3,448	Kawashima et al. 1998
A*6802	CLTSTVQLV	8,000	Kawashima et al. 1998
A*6802	YLQLVFGIEV	9,302	Kawashima et al. 1998
A*6802	LLTFWNPPV	9,442	Kawashima et al. 1998
A*6802	LLTFWNPPT	>10,000	Kawashima et al. 1998
A*6802	VMAGVGSPPV	>10,000	Kawashima et al. 1998
A*6802	ILHNGAYSL	>10,000	Kawashima et al. 1998
A*6802	YLSGANLNL	>10,000	Kawashima et al. 1998
A*6802	ALCRWGLLL	>10,000	Chang et al. 1999
B*3501	LPGCSFSIF	90.4	Chang et al. 1999
B*5101	LPGCSFSIF	100.0	Chang et al. 1999
B*5301	LPGCSFSIF	114.0	Chang et al. 1999
B*5401	LPGCSFSIF	6666.0	Chang et al. 1999

Table 2.2 Grouping of peptides based on IC_{50} binding values

Alleles	Peptides	GB	MB	NB
A*0201	23	15	8	–
A*0202	12	4	5	3
A*0203	13	11	2	–
A*0206	13	3	5	5
A*0301	8	7	1	–
A*1101	8	5	3	–
A*3101	8	2	1	5
A*3301	8	–	–	8
A*6801	8	1	5	2
A*6802	13	3	–	10
B*3501	1	1	–	–
B*5101	1	1	–	–
B*5301	1	–	1	–
B*5401	1	–	–	1
Total	118	53	31	34

GB good binder (<100 nM), *MB* moderate binder (100–500 nM), *NB* nonbinder/weak binder (>500 nM)

2.4 MHC–Peptide Structure Dataset Clustering

The MHC–peptide dataset was further clustered based on MHC allele specificity and peptide length. Each member of the clustered subgroup was analyzed for MHC–peptide interaction and discussed. The MHC–peptide structural data set was primarily clustered into class I MHC and class II MHC. Further, the class I MHC–peptide complexes were clustered into 13 subgroups based on the MHC alleles and the length of the bound peptide (Table 2.3). Similarly, the remaining eight MHC class II complexes were clustered into six subgroups (Table 2.4). The properties of MHC–peptide interactions identified from these 19 subgroups are summarized in Tables 2.5 and 2.6.

2.5 PDB Chain Identifier

The chain identifiers representing the alpha chain in all MHC class I entries except 1KBG (PDB code) is “a” (Table 2.3). “h” represents the chain identifier for the alpha chain in 1KBG. Most of the MHC class I-specific peptides are represented by the chain identifier “c,” while the rest are represented by the chain identifier “p” (Table 2.3). The “alpha” and “beta” chains in class II entries are represented by chain identifiers “a” and “b,” respectively (Table 2.4). The class II-specific peptides are represented by either one of the following chain identifiers: “c” or “e” or “b with some alphanumeric tags” (Table 2.4). An understanding of the current data

Table 2.3 Class I MHC-peptide data set

MHC source	Sub-groups	PDB code	MHC allele	CIM	Redundant peptide set	CIP	Nonredundant peptide set	PL	Peptide source	R(Å)	Release year
Human	Subset 1-group 1	1HHJ	A*0201	{a}	ILKEPVHGV	{c}	ILKEPVHGV	9	Synthetic	2.50	1993
Human		1AKJ	A*0201	{a}	Ilkepvhgv	{c}		9	HIV-1 RT	2.65	1997
Human		1HHK	A*0201	{a}	LLFGYPVYV	{c}	LLFGYPVYV	9	Synthetic	2.50	1993
Human		1AO7	A*0201	{a}	llfgypvyv	{c}		9	HTLV-1 Tax	2.60	1997
Human		1BD2	A*0201	{a}	llfgypvyv	{c}		9	HTLV-1 Tax	2.50	1998
Human		1B0G	A*0201	{a}	ALWGFPPVL	{c}	ALWGFPPVL	9	Human-peptide P1049	2.60	1998
Human		*1A9K	A*0201	{a}	alwgfppvl	{c}		9	Human-peptide P1049	2.50	1998
Human		1HHG	A*0201	{a}	TLTSCNTSV	{c}	TLTSCNTSV	9	HIV-1 gp 120	2.60	1993
Human		1HHI	A*0201	{a}	GILGFVFTL	{c}	GILGFVFTL	9	Synthetic	2.50	1993
Human		1B0R	A*0201	{a}	gilgfvtcde	{c}		9	Influenza matrix	2.90	1998
Human	Subset 1-group 2	2CLR	A*0201	{a}	MLLSVPLLLG	{c}	MLLSVPLLLG	10	Synthetic	2.00	1995
Human		1HHH	A*0201	{a}	FLPSDFPPSV	{c}	FLPSDFPPSV	10	HBV nucleocapsid	3.00	1993
Human	Subset 2-group 1	1TMC	A*6801	{a}	EVAPPEYHRK	{c}	EVAPPEYHRK	10	Synthetic	2.30	1995
Human	Subset 3-group 1	1AGB	B*0801	{a}	GGRKKYKL	{c}	GGRKKYKL	8	HIV-1 gag	2.20	1997
Human		1AGC	B*0801	{a}	GKKKKYQL	{c}	GKKKKYQL	8	HIV-1 gag	2.10	1997
Human		1AGD	B*0801	{a}	GKKKKYKL	{c}	GKKKKYKL	8	HIV-1 gag	2.05	1997
Human		1AGE	B*0801	{a}	GGRKKYRL	{c}	GGRKKYRL	8	HIV-1 gag	2.30	1997
Human		1AGF	B*0801	{a}	GKKKKYKL	{c}	GKKKKYKL	8	HIV-1 gag	2.20	1997
Human	Subset 4-group 1	1HSA	B*2705	{a}	ARAAAAAAA	{c}	ARAAAAAAA	9	-	2.10	1992

(continued)

Table 2.3 (continued)

MHC source	Sub-groups	PDB code	MHC allele	CIM	Redundant peptide set	CIP	Nonredundant peptide set	PL	Peptide source	R (Å)	Release year
Human	Subset	1A1N	B*3501	{a}	VPLRPMTY	{c}	VPLRPMTY	8	HIV-1 Nef	2.00	1998
5-group 1											
Human	Subset	1A9E	B*3501	{a}	LPPLDITPY	{c}	LPPLDITPY	9	EBV-Ebna3c	2.50	1998
5-group 2											
Human	Subset	1A9B	B*3501	{a}	lpplditpy	{c}	TPYDINQML	9	EBNA-3C	3.20	1998
Human		1A1M	B*5301	{a}	TPYDINQML	{c}		9	HIV-2 Gag	2.30	1998
6-group 1											
Human	Subset	1A1O	B*5301	{a}	KPIVQYDNF	{c}	KPIVQYDNF	9	HIV-1 Nef	2.30	1998
Murine		1OSZ	H-2KB	{a}	RGYLYQGL	{c}	RGYLYQGL	8	V _{sv} -nucleoprotein	2.10	1999
7-group 1											
Murine	Subset	2VAB	H-2KB	{a}	RGYVYQGL	{p}	RGYVYQGL	8	SV nucleoprotein	2.50	1996
Murine		1KBG	H-2KB	{h}	RGYV _Y uGL	{p}	SIINFEKL	8	Synthetic	2.20	1999
Murine		1VAC	H-2KB	{a}	SIINFEKL	{p}		8	Ovalbumin	2.50	1996
Murine		1VAD	H-2KB	{a}	SRDHSRTPM	{p}	SRDHSRTPM	9	Yeast α -glucosidase	2.50	1996
7-group 2											
Murine	Subset	2VAA	H-2KB	{a}	FAPGNYPAL	{p}	FAPGNYPAL	9	V _{sv} nucleoprotein	2.30	1996
Murine		1BZ9	H-2DB	{a}	FAPGVFPYM	{c}	FAPGVFPYM	9	Peptide P1027	2.80	1998
8-group 1											
Murine	Subset	1CE6	H-2DB	{a}	FAPGNYPAL	{c}	FAPGNYPAL	9	SV nucleoprotein	2.90	1999
Murine		1QLF	H-2DB	{a}	FAPSNYPAL	{c}	FAPSNYPAL	9	SV-nucleoprotein	2.65	1999
Murine		1BII	H-2DD	{a}	RGPGRAFVTI	{p}	RGPGRAFVTI	10	HIV-1 P ₁₈₋₁₁₀	2.40	1998
9-group 1											
Murine	Subset	1LDP	H-2LD	{a}	APAAAAAAM	{p}	APAAAAAAM	9	Natural peptide	3.10	1998
10-group 1											

PDB protein data bank, *PL* peptide length, *R* resolution, *CIM* chain identifier for MHC, *CIP* chain identifier for peptide, *Release year* the year in which the entry was released by PDB, obsolete entries are marked by “superscripted #” at the left end of the PDB code. Duplicate peptides were removed in the nonredundant peptide set

Table 2.4 Class II MHC-peptide dataset

MHC source	Sub groups	PDB code	MHC allele	CIM	Peptide	CIP	PL	Peptide source	R (Å)	Release year
Human	Subset 1-group 1	1AQD	DR1	{a,b}	*VGSDWRFRLRGYHQYA	{c}	15	Endogenous peptide	2.45	1998
Human	Subset 1-group 2	1DLH	DR1	{a,b}	PKYVKQNTLKLAT	{c}	13	Influenza virus	2.80	1994
Human		1SEB	DR1	{a,b}	AAAAAAAAAAAAA	{c}	13	Endogenous peptide	2.70	1996
Human	Subset 2-group 1	1BX2	DR2	{a,b}	ENPVVHFFKNIVTPR*	{c}	15	HMBP	2.60	1998
Human	Subset 3-group 1	1A6A	DR3	{a,b}	PVSKMRMATPLLMQA	{c}	15	CLIP fragment	2.75	1998
Human	Subset 4-group 1	2SEB	DR4	{a,b}	AYMRADAAAAGGA	{e}	12	Collagen II	2.50	1998
Murine	Subset 5-group 1	1IAO	I-A ^D	{a,b}	RGISQAVHAHAHEI	{b}	14	Egg ovalbumin	2.60	1998
Murine		2IAD	I-A ^D	{a,b}	GHATQGVTAASSHE	{b}	14	Influenza hemagglutinin	2.40	1998

PDB protein data bank, *PL* peptide length, *R* resolution, *CIM* chain identifier for MHC, *CIP* chain identifier for peptide, *Release year* the year in which the entry was released by PDB, an asterisk (*) marks those residues for which the structural information is not available

formats with reference to a specific biological function such as MHC-peptide binding will provide ways to curate them to a more consistent format and thus, aid in the development of an automated data retrieval system.

2.6 Information Redundancy in Dataset

In class I dataset, entries with PDB code 1AKJ, 1AO7, 1BD2, 1B0R, 1A9B, 1KBG are considered redundant by the authors (within the limitation of this chapter) as these entries are duplicates describing a particular sequence information, or they contain incomplete structural information. Specifically, the PDB entry 1A9K was classified as obsolete by PDB and hence it is indicated by a “#” (Table 2.3). In class II data set, the PDB entry 1AQD, represents a DR1–peptide complex, where the peptide sequence is *VGSDWRFLRGYHQYA. The coordinates for the first residue valine (V) are not available in the PDB file. Similarly, the peptide sequence in 1BX2 (PDB code) is ENPVVHFFKNIVTPR*, and coordinates for the last residue arginine (R) are not available. Residues for which the structural information is not available are marked by an asterisk mark (*) solely to indicate this feature (Table 2.4). In some entries, there are two complexes per asymmetric unit each composed of three polypeptide chains, describing identical MHC-peptide sequence data. In such cases, we take the data for one and leave the other.

2.7 Information from MHC-Peptide Data

Data of highest quality that is quantitatively rich in information content when extracting knowledge from data repositories are generally prepared (Rechenmann 2000). We show the use of 36 nonredundant MHC–peptide structural complexes from PDB for analysis. In the dataset, 28 of the 36 complexes were class I MHC–peptide complexes and the remaining 8 complexes were class II MHC–peptide complexes. Among 28 nonredundant class I entries listed (Table 2.3), 10 are murine H-2 peptide complexes and the remaining 18 are human MHC–peptide complexes. Similarly, of the eight class II entries listed (Table 2.4), two are murine complexes and the remaining six are human MHC–peptide molecules. The available structural information has tremendously improved our knowledge on peptide binding to MHC molecules. MHC–peptide complexes are available for HLA-A*0201, HLA-A*6801, HLA-B*0801, HLA-B*2701, HLA-B*3501, HLA-B*5301, H-2KB, H-2DB, H-2LD, DR1, DR2, DR3, DR4 and I-AD (Tables 2.3 and 2.4). Among them HLA-A*0201–peptide complexes are the most represented (Table 2.3). Therefore, representative structure-based binding information on allele-specific MHC peptide complexes will deduce the relationship that map the structure function differences influenced by sequence variation. The overwhelming advancements in r-DNA technology and high-throughput X-ray crystallography projects (Service 2000) will speed up MHC-peptide research in the near future.

2.8 Structural Parameters for MHC–Peptide Dataset Analysis

Interface area indicates a measure of the mean change in accessible area (mean Δ ASA) for the peptide and the MHC molecules when going from a monomeric MHC molecule to a dimeric MHC–peptide complex state (Tables 2.5 and 2.6). Solvent accessible surface area for the MHC–peptide complexes, monomeric peptides and monomeric MHC molecules are calculated using the algorithm implemented elsewhere (Lee and Richard 1971). The gap volume between the peptide and the MHC was calculated using SURFNET (Laskowski 1995). The number of intermolecular hydrogen bonds between the peptide and the MHC were calculated using HBPLUS (McDonald and Thornton 1994) in which hydrogen bonds are defined according to standard geometric criteria. Gap index, defining the complementarity of the interacting surfaces in the MHC-peptide have been evaluated by the formula as defined elsewhere (Jones and Thornton 1996). The entries for which the relative binding strengths are not easily available are marked by “-.”

2.9 Creation of Heterodimer and Homodimer Dataset

A total of 2,488 heterodimer (different subunits interacting) candidates and 1,324 homodimer (identical subunits interacting) candidates were downloaded from PDB and PQS (Protein Quaternary Structure) server. We then created a nonredundant dataset of 156 heterodimers (Table 2.7) and 170 homodimers (Table 2.8) such that they satisfy the following conditions. These include: (1) each chain ≥ 50 residues; (2) structures determined by X-ray crystallography; (3) resolution ≤ 2.5 Å; (4) the structure with the highest resolution was selected where more than one structure was available; (5) redundant entries were removed at a sequence similarity cutoff of $\geq 30\%$.

2.10 Homodimer Folding Dataset

We created a dataset consisting of 41 homodimer complex structures (2S (two state): 25; 3SDI (three state with dimer intermediate): 5; and 3SMI (three state with monomer intermediate): 10 from PDB). The unfolding pathways for these dimers observed using thermodynamic experiments that were obtained from literature (Table 2.9). The selected homodimers are at least 40 residues per monomer.

2.11 Alanine-Mutated Interface Residues Dataset

A dataset of 296 alanine-mutated interface residues (Table 2.10) derived from 15 protein–protein complexes (Table 2.11) was obtained from ASEdb (Alanine Scanning Energetics database). [26] These residues have $\Delta\Delta G$ in the range -0.9 – 10

Table 2.5 Class I MHC–peptide interface (nonredundant set)

PDB code	Subgroups	Binding strength	Interface area (Å ²)	H-bonds [number]	Volume (Å ³)	Gap index (Å)
1HHJ	Subset 1-group 1	242 ^a [IC ₅₀]	880.4	14	827.4	0.94
1HHK		2.5 ^b [IC ₅₀]	885.0	10	1083.4	1.22
1BOG		–	869.0	12	441.3	0.51
1HHG		–	803.3	12	1039.9	1.29
1HHI		6 ^c [IC ₅₀]	857.8	9	455.7	0.53
2CLR	Subset 1-group 2	–	910.9	10	911.3	1.00
1HHH		2.5 ^d [IC ₅₀]	940.5	11	655.9	0.70
1TMC	Subset 2-group 1	–	955.5	14	926.2	0.97
1AGB	Subset 3-group 1	–	844.4	15	881.7	1.04
1AGC		–	830.5	18	688.1	0.83
1AGD		–	846.0	16	816.1	0.96
1AGE		–	832.8	15	920.6	1.10
1AGF		–	883.1	14	765.4	0.87
1HSA	Subset 4-group 1	–	727.8	14	1148.4	1.58
1A1N	Subset 5-group 1	–	879.7	11	670.2	0.76
1A9E	Subset 5-group 2	–	895.5	12	779.3	0.87
1A1M	Subset 6-group 1	–	845.4	12	971.2	1.14
1A1O		–	994.5	10	778.8	0.78
1OSZ		–	946.9	18	756.2	0.80
2VAB	Subset 7-group 1	–	881.7	12	1301.0	1.47
1VAC		5900 ^e [M ⁻¹ s ⁻¹]	892.2	14	691.2	0.77
1VAD		–	880.5	21	939.5	1.07
2VAA		–	938.2	16	738.3	0.79
1BZ9	Subset 8-group 1	–	884.0	10	897.0	1.01
1CE6		–	867.7	15	787.7	0.91
1QLF		–	893.7	13	567.3	0.63
1BII	Subset 9-group 1	–	937.4	14	792.0	0.84
1LDP	Subset 10-group 1	–	771.4	9	889.3	1.15

Interface area indicates a measure of the mean change in accessible area (mean Δ ASA) for the peptide and the MHC molecules when going from a monomeric MHC molecule to a dimeric MHC–peptide complex state. Solvent accessible surface area both for the MHC–peptide complexes as well as the individual peptides and MHC molecules was calculated using the algorithm implemented by Lee and Richard (1971). The gap volume between the peptide and the MHC was calculated using SURFNET (Laskowski 1995). The number of intermolecular hydrogen bonds between the peptide and the MHC were calculated using HBPLUS (McDonald and Thornton 1994) in which hydrogen bonds are defined according to standard geometric criteria. Gap index, defining the complementarity of the interacting surfaces in the MHC–peptide have been evaluated by the formula as defined elsewhere (Jones and Thornton 1996). The derived knowledge presented here is only for the nonredundant peptide set. Binding strength for these peptides was expressed in IC₅₀ units, denoting the peptide’s concentration required to inhibit the binding of the standard peptide by 50%. The entries for which the relative binding strengths are not easily available are marked by “–”. References: ^aSette et al. (1994), ^bLauvau et al. (1999), ^cGianfrani et al. (2000), ^dLivingston et al. (1999), ^eChen et al. (1994)

Table 2.6 Class II MHC-peptide interface (nonredundant set)

PDB code	Sub-groups	Binding strength	Interface area (\AA^2)	H-bonds (number) (total)	Chain A and peptide	Chain B and peptide	Volume (\AA^3)	Gap index (\AA)
1AQD	Subset 1-group 1	–	1211.2	18	10	8	1182.7	0.98
1DLH	Subset 1-group 2	–	1168.4	17	10	7	1081.8	0.93
1SEB		–	834.8	12	7	5	964.1	1.15
1BX2	Subset 2-group 1	–	1034.4	15	9	6	1308.6	1.27
1A6A	Subset 3-group 1	–	1171.0	19	10	9	1204.7	1.03
2SEB	Subset 4-group 1	–	960.5	14	5	9	836.1	0.87
1IAO	Subset 5-group 1	–	1087.4	14	8	6	1449.5	1.33
2IAD	Subset 5-group 2	–	963.5	13	7	6	1475.1	1.53

Interface area indicates a measure of the mean change in accessible area (mean ΔASA) for the peptide and the MHC molecules when going from a monomeric MHC molecule to a dimeric MHC-peptide complex state. Solvent accessible surface area for the MHC-peptide complexes, monomeric peptides and monomeric MHC molecules is calculated using the algorithm implemented elsewhere (Lee and Richard 1971). The gap volume between the peptide and the MHC was calculated using SURFNET (Laskowski 1995). The number of intermolecular hydrogen bonds between the peptide and the MHC were calculated using HBPLUS (McDonald and Thornton 1994) in which hydrogen bonds are defined according to standard geometric criteria. Gap index, defining the complementarity of the interacting surfaces in the MHCp have been evaluated by the formula as defined elsewhere (Jones and Thornton 1996). The entries for which the relative binding strengths are not easily available are marked by “–”.

Table 2.7 Heterodimer dataset

PDB code	Resolution (Å)	Chain one	Name of chain one	Length	Chain two	Name of chain two	Length
1YCS	2.2	B	53BP2	193	A	P53	191
1ABR	2.1	B	Abrin-A	267	A	Carbohydrate	251
1KU6	2.5	A	Acetylcholinesterase	535	B	Fasciculin 2	61
1LFD	2.1	B	Active ras protein	167	A	Ras-interacting domain of ralgs	87
1JIW	1.7	P	Alkaline metalloproteinase	470	I	Proteinase inhibitor	105
1BPL	2.2	B	Alpha-amylase	290	A	Alpha-amylase	179
1KXV	1.6	A	Alpha-amylase	496	C	Camelid VHH domain	119
1TMQ	2.5	A	Alpha-amylase	470	B	Ragi bifunctional inhibitor	117
1BVN	2.5	P	Alpha-amylase	496	T	Tendamistat	71
1ACB	2.0	E	Alpha-chymotrypsin	241	I	Eglin C	63
1CHO	1.8	E	Alpha-chymotrypsin	238	I	Turkey ovomucoid third domain	53
1CGI	2.3	E	Alpha-chymotrypsinogen	245	I	Trypsin inhibitor	56
1SLU	1.8	B	Anionic trypsin	216	A	Ecotin	131
1RE0	2.4	B	ARF guanine-nucleotide exchange factor 1	195	A	ADP-ribosylation factor 1	162
1KSH	1.8	A	ARF-like protein 2	164	B	Cyclic phosphodiesterase delta-subunit	141
1MG9	2.3	B	ATP dependent CLP protease	143	A	Protein YLJA	84
1BRL	2.4	A	Bacterial luciferase	340	B	Bacterial luciferase	319
1AVA	1.9	A	Barley alpha-amylase 2	403	C	Barley alpha-amylase/subtilisin inhibitor	181
1B27	2.1	A	Barnase	110	D	Barstar	90
1LUJ	2.5	A	Beta-catenin	501	B	Beta-catenin-interacting protein ICAT	71
1S0W	2.3	A	Beta-lactamase tem	263	C	Beta-lactamase inhibitory protein	165

1BND	2.3	A	Brain derived neurotrophic factor	109	B	Neurotrophin 3	108
1D4X	1.8	A	C. Elegans actin 1/3	368	G	Gelsolin	124
1G4Y	1.6	R	Calmodulin	147	B	Calcium-activated potassium channel RSK2	81
1DTD	1.7	A	Carboxypeptidase A2	303	B	Metallocarboxypeptidase inhibitor	61
1NW9	2.4	B	Catalytic domain of caspase-9	238	A	Inhibitor of apoptosis protein 3	91
1OKK	2.1	D	Cell division protein	265	A	Signal recognition particle protein	290
1H1S	2.0	A	Cell division protein kinase 2	296	B	Cyclin A2	258
1OHZ	2.2	A	Cellulosomal scaffolding protein A	140	B	Endo-1,4-beta-xy lanase Y	56
1HL6	2.5	A	CG8781 protein	119	B	Mago nashi protein	137
1P5V	1.7	A	Chaperone protein CAFIM	191	B	F1 capsule antigen	136
1PDK	2.4	A	Chaperone protein PAPD	296	B	Protein PAPP	258
1N0L	2.3	A	Chaperone protein PAPD	212	B	Mature fimbrial protein PAPE	116
1FFG	2.1	B	Chemotaxis protein chea	68	A	Chemotaxis protein chey	128
1EAY	2	A	Chey	128	C	Chea	67
1P2M	1.8	A	Chymotrypsinogen A	238	B	Pancreatic trypsin inhibitor	58
1HCG	2.2	A	Coagulation factor	236	B	Coagulation factor	51
1V74	2.0	A	Colicin D	107	B	Colicin D immunity protein	87
1E44	2.4	B	Colicin E3	96	A	Immunity protein	84
1FR2	1.6	B	Colicin E9	131	A	Colicin E9 immunity protein	83
1F5Q	2.5	A	Cyclin dependent kinase 2	296	B	Gamma herpesvirus cyclin	247
1FIN	2.3	A	Cyclin-dependent kinase	298	B	Cyclin A	260

(continued)

Table 2.7 (continued)

PDB code	Resolution (Å)	Chain one	Name of chain one	Length	Chain two	Name of chain two	Length
IBLX	1.9	A	Cyclin-dependent kinase 6	305	B	P19ink4D	160
IM9E	1.7	A	Cyclophilin A	164	D	HIV-1 capsid	135
IS6V	1.9	A	Cytochrome C peroxidase	294	B	Cytochrome C	108
IR8S	1.5	E	Cytohesin 2	187	A	ADP-ribosylation factor 1	160
1UJZ	2.1	B	Designed colicin E7 dnase	127	A	Designed colicin E7 immunity protein	87
INLV	1.8	A	Dictyostelium discoideum actin	364	G	Gelsolin	123
1H31	1.5	A	Diheme cytochrome C	260	B	Cytochrome C	138
1EM8	2.1	A	DNA polymerase III CHI subunit	147	B	DNA polymerase III PSI subunit	110
1JQL	2.5	A	DNA polymerase III, beta chain	366	B	DNA polymerase III delta subunit	140
1EAI	2.4	A	Elastase	240	C	Chymotrypsin isoinhibitor I	61
1EFV	2.1	A	Electron transfer flavoprotein alpha chain	312	B	Electron transfer flavoprotein beta chain	252
1F60	1.7	A	Elongation factor EEF1A	440	B	Elongation factor EEF1BA	90
1TA3	1.7	B	Endo-1,4-beta-xylanase	301	A	Xylanase inhibitor protein I	274
1TE1	2.5	B	Endo-1,4-xylanase	190	A	Xylanase inhibitor protein I	274
3FAP	1.9	A	FKBP12-binding protein	107	B	FKBP12-rapamycin associated protein	94
1FCD	2.5	A	Flavocytochrome C sulfide dehydrogenase	401	C	Flavocytochrome c sulfide dehydrogenase	174
1NF3	2.1	A	G25k GTP-binding protein	194	C	PAR-6B	123
1NQI	2	B	Galactosyltransferase	272	A	Alpha lactalbumin	123
1WQ1	2.5	G	Gapette	320	R	Harvey-RAS	166

1OR0	2.0	B	Glutaryl acylase beat subunit	510	A	Glutaryl acylase alpha subunit	152
1AXI	2.1	B	Growth hormone receptor	191	A	Growth hormone	175
2NGR	1.9	B	Gipase activating protein	196	A	GTP-binding protein	191
1TX4	1.7	A	Gipase-activating protein rho gap	196	B	Transforming protein RHOA	174
1AY7	1.7	A	GuanyI-specific ribonuclease SA	96	B	Barstar	89
1HX1	1.9	A	Heat shock cognate 71 KDA	377	B	Bag-family molecular chaperone regulator-1	112
1USU	2.2	A	Heat shock protein HSP82	246	B	AHA1	132
2HBE	2.0	B	Hemoglobin	146	A	Hemoglobin	141
1GPW	2.4	A	Hisf protein	253	B	Amidotransferase HISF	200
1CXZ	2.2	A	His-tagged transforming protein RHOA	182	B	PKN	86
1US7	2.3	B	HSP90 chaperone protein kinase	194	A	Heat shock protein HSP82	207
1KXP	2.1	D	Human vitamin D-binding protein	438	A	Actin, alpha skeletal muscle	349
1H2A	1.8	L	Hydrogenase	534	S	Hydrogenase	267
1KA9	2.3	F	Imidazole glycerol phosphatase synthase	251	H	Imidazole glycerol phosphatase synthase	195
1IBR	2.3	B	Importin beta-1 subunit	458	A	GTP-binding nuclear protein ran	169
1PVH	2.5	A	Interleukin 6 signal transducer	201	B	Leukemia inhibitory factor	169
1IAR	2.3	B	Interleukin-4 receptor alpha chain	188	A	Interleukin	129
1IIR	2.4	A	Interleukin-6 receptor beta chain	301	B	Viral IL-6	167
1O6S	1.8	A	Internalin A	461	B	E-cadherin	105

(continued)

Table 2.7 (continued)

PDB code	Resolution (Å)	Chain one	Name of chain one	Length	Chain two	Name of chain two	Length
1K1I	2.3	B	Intersectin long form	342	A	G25k GTP-binding protein	178
2KIN	1.9	A	Kinesin	238	B	Kinesin	100
1PPF	1.8	E	Leukocyte elastase	218	I	Ovomucoid inhibitor	56
1OP9	1.9	B	Lysozyme C	130	A	H16 camel VHH fragment	121
1UUZ	1.8	D	Lysozyme C	129	A	Inhibitor of vertebrate lysozyme	130
1O00	1.9	A	Mago nashi protein	144	B	Drosophila Y14	92
1SVX	2.2	B	Maltose-binding periplasmic protein	369	A	Ankyrin repeat protein OFF7	157
1PQZ	2.1	A	MCMV M144	238	B	Beta-2-microglobulin	99
1MEE	2.0	A	Mesentericopeptidase	275	I	Eglin-C	64
1JW9	1.7	B	Molybdopterin biosynthesis moeb protein	240	D	Molybdopterin converting factor	81
1Q40	2.0	B	Mma export factor MEX67	180	A	Mrna transport regulator MTR2	163
1SHW	2.2	B	Neural kinase	181	A	Ephrin-A5	138
1QAV	1.9	B	Neuronal nitric oxide synthase	115	A	Alpha-1 syntrophin	90
1E96	2.4	B	Neutrophil cytosol factor 2	185	A	Ras-related C3 botulinum toxin substrate 1	178
1NPE	2.3	A	Nidogen	263	B	Laminin gamma-1 chain	164
1GL4	2.0	A	Nidogen-1	273	B	Proteoglycan core protein	89
1M4U	2.4	A	Noggin	199	L	Osteogenic protein 1	112
1FYH	2.0	A	Nterferon-gamma	242	B	Interferon-gamma receptor alpha chain	201
1STF	2.4	E	Papain	212	I	Stefin B	98
1F34	2.5	A	Pepsin A	325	B	Major pepsin inhibitor PI-3	138
1UBK	1.2	L	Periplasmic hydrogenase large subunit	534	S	Periplasmic hydrogenase small subunit	267

1JLT	1.4	B	Phospholipase A2	122	A	Phospholipase A2 inhibitor	122
1L4Z	2.3	A	Plasminogen	248	B	Streptokinase	125
1DHK	1.9	A	Porcine pancreatic alpha-amylase	495	B	Bean lectin-like inhibitor	195
3YGS	2.5	P	Procaspase 9	97	C	Apoptotic protease activating factor 1	95
1FT1	2.3	B	Protein farnesyltransferase	416	A	Protein farnesyltransferase	315
1G4U	2.3	S	Protein tyrosine phosphatase SPTP	360	R	Ras-related C3 botulinum toxin substrate 1	180
1CT4	1.6	E	Proteinase	185	I	Ovomucoid inhibitor	51
1VG0	2.2	A	Rab escort protein 1	481	B	Ras-related protein rab-7	182
1F2T	1.6	A	Rad50 abc-atpase N-terminal domain	145	B	Rad50 abc-atpase C-terminal domain	143
1GUA	2.0	A	Rap1A	167	B	C-raf1	76
1HE1	2.0	C	Ras-related C3 botulinum toxin substrate 1	176	A	Exoenzyme S	135
1DS6	2.4	A	Ras-related C3 botulinum toxin substrate 2	181	B	RHO GDP-dissociation inhibitor 2	179
1C1Y	1.9	A	Ras-related protein	167	B	Proto-onkogene serine	77
1DFJ	2.5	E	Ribonuclease A	124	I	Ribonuclease inhibitor	456
1DZB	2.0	A	SCFV fragment 1F9	224	X	Turkey egg-white lysozyme C	129
1H2S	1.9	A	Sensory rhodopsin II	225	B	Sensory rhodopsin II transducer	60
1P57	1.8	B	Serine protease hepsin heavy chain	247	A	Serine protease hepsin light chain	110
4SGB	2.1	E	Serine proteinase B	185	I	Potato inhibitor	51
1SMP	2.3	A	Serratia metallo proteinase	468	I	Erwinia chrysanthemi inhibitor	100

(continued)

Table 2.7 (continued)

PDB code	Resolution (Å)	Chain one	Name of chain one	Length	Chain two	Name of chain two	Length
1NRJ	1.7	B	Signal recognition particle receptor	191	A	Docking protein	147
1RJ9	1.9	A	Signal recognition protein	277	B	Signal recognition particle protein	282
1JTP	1.9	A	Single-domain antibody	135	L	Lysozyme C	129
1SGD	1.8	E	Streptogrisin B	185	I	Ovomucoid	51
1LW6	1.5	E	Subtilisin BPN	281	I	Ubtilisin-chymotrypsin inhibitor-2A	63
2SIC	1.8	E	Subtilisin BPN	275	I	Streptomyces subtilisin inhibitor	107
1SPB	2.0	S	Subtilisin BPN	264	P	Subtilisin BPN prosegment	71
1R0R	1.1	E	Subtilisin carlsberg	274	I	Ovomucoid	51
1CSE	1.2	E	Subtilisin carlsberg	274	I	Eglin-C	63
1SCJ	2.0	A	Subtilisin E	275	B	Subtilisin E	71
2SNI	2.1	E	Subtilisin novo	275	I	Chymotrypsin inhibitor 2	64
1EUC	2.1	B	Succinyl-coa synthetase, beta chain	393	A	Succinyl-coa synthetase, alpha chain	306
1ONQ	2.2	A	T-cell surface glycoprotein CD1A	274	B	Beta-2-microglobulin	99
1JTD	2.3	A	Tem-1 beta-lactamase	262	B	Beta-lactamase inhibitor protein II	273
1KTZ	2.2	B	TGF-beta type II receptor	106	A	Transforming growth factor beta 3	82
2TEC	2.0	E	Thermitase	279	I	Eglin-C	63
1JKG	1.9	B	Tip associating protein	180	A	NTF2-related export protein 1	139
1D4V	2.2	B	TNF-related apoptosis inducing ligand	163	A	Death receptor 5	117
1AVW	1.8	A	Trypsin	223	B	Trypsin inhibitor	171

1BRB	2.1	E	Trypsin	223	I	BPTI	51
1F5R	1.7	A	Trypsin II	216	I	Pancreatic trypsin inhibitor	57
1K9O	2.3	E	Trypsin II anionic	223	I	Alaserpin	376
1D6R	2.3	A	Trypsinogen	223	I	Bowman-birk proteinase inhibitor	58
1OPH	2.3	B	Trypsinogen	223	A	Alpha-1 protease inhibitor	375
1P2J	1.4	A	Trypsinogen	220	I	Pancreatic trypsin inhibitor	56
1S1Q	2.0	A	Tumor susceptibility gene 101 protein	137	B	Ubiquitin	71
1ITB	2.5	B	Type 1 interleukin-1 receptor	310	A	Interleukin-1 beta	153
1J7D	1.9	B	Ubiquitin-conjugating enzyme E2-17 KDA	149	A	MMS2	140
1EUV	1.3	A	ULP1 protease	221	B	Ubiquitin-like protein SMT3	79
1UGH	1.9	E	Uracil-dna glycosylase	223	I	Uracil-DNA glycosylase inhibitor	82
1UZX	1.9	A	Vacuolar protein sorting-associated protein	135	B	Ubiquitin	75
1JTT	2.1	A	VH single-domain antibody	133	L	Lysozyme	129
1RKE	2.4	A	Vinculin	262	B	VCL protein	176
1MA9	2.4	A	Vitamin D-binding protein	442	B	Actin, alpha skeletal muscle	356
1YVN	2.1	A	Yeast actin	372	G	Gelsolin	125
1OXB	2.3	A	YDP1P	166	B	Osmolarity two-component system protein	124

Table 2.8 Homodimer dataset

PDB	Resolution (Å)	Name	Source	Chain one	Length	Chain two	Length
1CNZ	1.8	3-isopropylmalate dehydrogenase	<i>Salmonella typhimurium</i>	A	363	B	363
1AFW	1.8	3-ketoacyl-coa thiolase	<i>Saccharomyces cerevisiae</i>	A	390	B	393
1M4I	1.5	Acetyltransferase	<i>Escherichia coli</i>	A	181	B	176
1LQ9	1.3	Actva-orf6 monooxygenase	<i>Streptomyces coelicolor</i>	A	112	B	112
1ADE	2	Adenylosuccinate synthetase	<i>E. coli</i>	A	431	B	431
1M7H	2	Adenylylsulfate kinase	<i>Penicillium chrysogenum</i>	A	203	B	200
1NA8	2.3	ADP-ribosylation-binding protein	<i>Homo sapiens</i>	A	151	B	145
1OR4	2.2	Aerotactic transducer hemat	<i>Bacillus subtilis</i>	A	169	B	158
1BD0	1.6	Alanine racemase	<i>Bacillus stearothermophilus</i>	A	381	B	380
1A4U	1.9	Alcohol dehydrogenase	<i>Drosophila lebanonensis</i>	A	254	B	254
1ALK	2	Alkaline phosphatase	<i>E. coli</i>	A	449	B	449
1LK9	1.5	Alliin lyase	<i>Allium sativum</i>	A	425	B	427
1HSS	2.1	Alpha-amylase inhibitor	<i>Triticum aestivum</i>	A	111	B	111
1S2Q	2.1	Amine oxidase B	<i>H. sapiens</i>	A	499	B	494
1EKP	2.5	Amino acid aminotransferase	<i>H. sapiens</i>	A	365	B	365
2GSA	2.4	Aminotransferase	<i>Synechococcus sp.</i>	A	427	B	427
1DQT	2	Antigen	<i>Mus musculus</i>	A	117	B	117
1BJW	1.8	Aspartate aminotransferase	<i>Thermus thermophilus</i>	A	381	B	381
1JFL	1.9	Aspartate racemase	<i>E. coli</i>	A	228	B	228
1MJH	1.7	Atp-binding protein	<i>Methanococcus jannaschii</i>	A	143	B	144
1IRI	2.4	Autocrine motility factor	<i>H. sapiens</i>	A	557	B	557
1LR5	1.9	Auxin-binding protein	<i>Z. mays</i>	A	160	B	160
1N80	2.5	Baseplate structural protein	Bacteriophage T4	A	328	B	328
1EWZ	2.4	Beta lactamase oxa-10	<i>Pseudomonas aeruginosa</i>	A	243	C	243
1EBL	1.8	Beta-ketoacyl-acp Synthase III	<i>E. coli</i>	A	309	B	309
1N1B	2	Bornyl diphosphate synthase	<i>Salvia officinalis</i>	A	516	B	519

1KSO	1.7	Calcium-binding protein A3	<i>H. sapiens</i>	A	93	B	93
1JD0	1.5	Carbonic anhydrase	<i>H. sapiens</i>	A	260	B	259
1AUO	1.8	Carboxylesterase	<i>Pseudomonas fluorescens</i>	A	218	B	218
1CDC	2	CD2	<i>Rattus norvegicus</i>	A	96	B	96
1F13	2.1	Cellular coagulation factor	<i>H. sapiens</i>	A	722	B	719
1NW1	2	Choline kinase	<i>Caenorhabditis elegans</i>	A	365	B	357
1R5P	2.2	Circadian oscillation regulator	<i>Anabaena</i> sp.	A	90	B	93
1G64	2.1	Cob(I) alamin adenosyltransferase	<i>S. typhimurium</i>	A	169	B	190
1OTV	2.1	Coenzyme pqq synthesis protein C	<i>Klebsiella pneumoniae</i>	A	254	B	254
1I0R	1.5	Conserved hypothetical protein	<i>Archaeoglobus fulgidus</i>	A	161	B	168
1OAC	2	Copper amine oxidase	<i>E. coli</i>	A	719	B	722
1EAJ	1.4	Coxsackie virus	<i>H. sapiens</i>	A	124	B	120
1CHM	1.9	Creatinase	<i>Pseudomonas putida</i>	A	401	B	401
1S44	1.6	Crustacyanin A1 subunit	<i>Homarus gammarus</i>	A	180	B	180
1GD7	2	CSAA protein	<i>T. thermophilus</i>	A	109	B	109
1L5B	2	Cyanovirin-N	<i>Nostoc ellipsosporum</i>	A	101	B	101
1SO2	2.4	Cyclic Phosphodiesterase B	<i>H. sapiens</i>	A	363	B	363
1P3W	2.1	Cysteine desulfurase	<i>E. coli</i>	A	385	B	385
1COZ	2	Cytidyltransferase	<i>B. subtilis</i>	A	126	B	126
1P6O	1.1	Cytosine deaminase	<i>S. cerevisiae</i>	A	156	B	161
2DAB	2	D-amino acid aminotransferase	<i>Thermophilic bacillus</i>	A	280	B	282
1F17	2.3	Dehydrogenase	<i>H. sapiens</i>	A	293	B	291
2NAC	1.8	Dehydrogenase	<i>Methylotrophic bacterium pseudomonas</i>	A	374	B	374
1NFZ	2	Delta-isomerase	<i>E. coli</i>	A	176	B	180
1D1G	2.1	Dihydrofolate reductase	<i>Thermotoga maritima</i>	A	164	B	164

(continued)

Table 2.8 (continued)

PDB	Resolution (Å)	Name	Source	Chain one	Length	Chain two	Length
IDOR	2	Dihydroorotate dehydrogenase A	<i>Lactococcus lactis</i>	A	311	B	311
IAD1	2.2	Dihydropteroate synthetase	<i>Staphylococcus aureus</i>	A	264	B	251
INU6	2.1	Dipeptidyl peptidase	<i>H. sapiens</i>	A	728	B	728
IPE0	1.7	DJ-1	<i>H. sapiens</i>	A	187	B	187
IG1A	2.5	DTDP-D-glucose 4,6-Dehydratase	<i>Salmonella enterica</i>	A	352	B	352
IBBH	1.8	Electron transport	<i>Chromatium vinosum</i>	A	131	B	131
IQ8R	1.9	Endodeoxyribonuclease rusa	<i>E. coli</i>	A	118	B	109
IRVE	2.5	Endonuclease	<i>E. coli</i>	A	244	B	244
IM9K	2	Endothelial nitric-oxide synthase	<i>H. sapiens</i>	A	400	B	401
IP43	1.8	Enolase 1	<i>S. cerevisiae</i>	A	436	B	436
IJR8	1.5	Erv2 protein mitochondrial	<i>S. cerevisiae</i>	A	105	B	105
IV26	2.5	Fatty-acid-coa synthetase	<i>T. thermophilus</i>	A	489	B	510
ILBQ	2.4	Ferrochelatase	<i>S. cerevisiae</i>	A	356	B	354
IRYA	1.3	Gdp-mannose mannosyl hydrolase	<i>E. coli</i>	A	160	B	160
IQFH	2.2	Gelation factor	<i>Dictyostelium discoideum</i>	A	212	B	212
IJV3	2.2	Glnac1p uridylyltransferase	<i>H. sapiens</i>	A	490	B	484
IDPG	2	Glucose 6-phosphate dehydrogenase	<i>Leuconostoc mesenteroides</i>	A	485	B	485
IQXR	1.7	Glucose-6-phosphate isomerase	<i>Pyrococcus furiosus</i>	A	187	B	187
IEOG	2.1	Glutathione S-transferase	<i>E. coli</i>	A	208	B	208
IN2A	1.9	Glutathione S-transferase	<i>E. coli</i>	A	201	B	187
IM0W	1.8	Glutathione synthetase	<i>S. cerevisiae</i>	A	481	B	479
IR9C	1.8	Glutathione transferase	<i>Mesorhizobium loti</i>	A	125	B	118
IF4Q	1.9	Grancalcin	<i>H. sapiens</i>	A	161	B	165

1DQP	1.8	Guanine phosphoribosyltransferase	<i>Giardia lamblia</i>	A	230	B	230
3SDH	1.4	Hemoglobin	<i>Scapharca inaequivalvis</i>	A	145	B	145
1IPI	2.2	Holliday junction resolvase	<i>P. furiosus</i>	A	114	B	114
1FWL	2.3	Homoserine kinase	<i>M. jannaschii</i>	A	296	B	296
2HHM	2.1	Hydrolase	<i>H. sapiens</i>	A	272	B	272
1PP2	2.5	Hydrolase	<i>Crotalus atrox</i>	R	122	L	122
1FJH	1.7	Hydroxysteroid dehydrogenase	<i>Comamonas testosteroni</i>	A	236	B	236
1GOS	1.9	Hypothetical protein	<i>E. coli</i>	A	201	B	202
1JOG	2.4	Hypothetical protein	<i>Haemophilus influenzae</i>	A	129	B	129
1PT5	2	Hypothetical protein	<i>E. coli</i>	A	415	B	415
1QYA	2	Hypothetical protein	<i>E. coli</i>	A	293	B	307
1FUX	1.8	Hypothetical protein	<i>E. coli</i>	A	164	B	163
1J30	1.7	Hypothetical rubrerythrin	<i>Sulfolobus tokodaii</i>	A	141	B	137
1LHZ	2.3	Immunoglobulin lambda	<i>H. sapiens</i>	A	213	B	213
1AA7	2.1	Influenza virus matrix protein	Influenza virus	A	158	B	157
8PRK	1.9	Inorganic pyrophosphatase	<i>S. cerevisiae</i>	A	282	B	282
1R8J	2	Kaia	<i>Synechococcus elongatus</i>	A	272	B	264
1CQS	1.9	Ketosteroid isomerase	<i>P. putida</i>	A	124	B	124
1AQ6	2	L-2-haloacid dehalogenase	<i>Xanthobacter autotrophicus</i>	A	245	B	245
1I2W	1.7	Lactamase	<i>Bacillus licheniformis</i>	A	255	B	256
1BH5	2.2	Lactoylgutathione lyase	<i>H. sapiens</i>	A	177	B	182
1QMJ	2.2	Lectin	<i>Gallus gallus</i>	A	132	B	132
1K75	1.8	L-histidinol dehydrogenase	<i>E. coli</i>	A	425	B	425
1EHI	2.4	Ligase	<i>L. mesenteroides</i>	A	360	B	347
1NWW	1.2	Limonene-1,2-epoxide hydrolase	<i>Rhodococcus erythropolis</i>	A	145	B	146
1UC8	2	Lysine biosynthesis enzyme	<i>T. thermophilus</i>	A	240	B	239

(continued)

Table 2.8 (continued)

PDB	Resolution (Å)	Name	Source	Chain one	Length	Chain two	Length
1EN5	2.3	Manganese superoxide dismutase	<i>E. coli</i>	A	205	B	205
1A4I	1.5	Methylenetetrahydrofolate	<i>H. sapiens</i>	A	285	B	295
1FC5	2.2	Molybdopterin biosynthesis	<i>E. coli</i>	A	397	B	396
1JYS	1.9	Mta/sah nucleosidase	<i>E. coli</i>	A	226	B	226
1LNW	2.1	Multidrug resistance operon repressor	<i>P. aeruginosa</i>	A	137	B	135
1FP3	2	N-acyl-d-glucosamine	<i>Sus scrofa</i>	A	402	B	402
1FYD	2.3	NAD(+) Synthetase	<i>B. subtilis</i>	A	271	B	246
1HJ3	1.6	Nitrite reductase	<i>Paracoccus pantotrophus</i>	A	544	B	542
1GIM	2.3	Nitrogenase iron protein	<i>Azotobacter vinelandii</i>	A	287	B	289
1G8T	1.1	Nuclease SM2 isoform	<i>Serratia marcescens</i>	A	241	B	241
1EYV	1.6	N-utilizing substance protein	<i>Mycobacterium tuberculosis</i>	A	131	B	133
1M98	2.1	Orange carotenoid protein	<i>Arthrospira maxima</i>	A	316	B	314
1ORO	2.4	Orotate	<i>E. coli</i>	A	213	B	206
1DVJ	1.5	phosphoribosyltransferase Orotidine 5(-phosphate decarboxylase	<i>Methanobacterium thermoautotrophicum</i>	A	239	B	211
1GGQ	2.5	Outer surface protein C	<i>Borrelia burgdorferi</i>	A	162	B	162
1AOR	2.3	Oxidoreductase	<i>P. furiosus</i>	A	605	B	605
1BMD	1.9	Oxidoreductase	<i>Thermus flavus</i>	A	327	B	327
1HDY	2.5	Oxidoreductase	<i>H. sapiens</i>	A	374	B	374
1N2O	2.1	Pantothenate synthetase	<i>M. tuberculosis</i>	A	279	B	279
1RN5	2.2	Peptide deformylase	<i>Leptospira interrogans</i>	A	177	B	177
1PN2	2	Peroxisomal hydratase	<i>Candida tropicalis</i>	A	269	B	267
1PN0	1.7	Phenol 2-monoxygenase	<i>Trichosporon cutaneum</i>	A	652	C	656
1BXG	2.3	Phenylalanine dehydrogenase	<i>Rhodococcus</i> sp.	A	349	B	347
1M6P	1.8	Phosphate receptor	<i>Bos Taurus</i>	A	146	B	146

PDB	Resolution (Å)	Name	Source	Chain one	Length	Chain two	Length
IRQL	2.4	Phosphonoacetaldehyde hydrolase	<i>Bacillus cereus</i>	A	257	B	257
IO4U	2.5	Phosphoribosyltransferase	<i>T. maritima</i>	A	265	B	266
IEZ2	1.9	Phosphotriesterase	<i>Pseudomonas diminuta</i>	A	328	B	328
IEXQ	1.6	Pol polyprotein	<i>E. coli</i>	A	147	B	145
IMNA	1.8	Polyketide synthase	<i>Streptomyces venezuelae</i>	A	276	B	278
IC6X	2.5	Protease	<i>E. coli</i>	A	99	B	99
IFL1	2.2	Protease	<i>E. coli</i>	A	192	B	207
IF89	2.4	Protein YLC351C	<i>S. cerevisiae</i>	A	271	B	271
ILHP	2.1	Pyridoxal kinase	<i>Ovis aries</i>	A	306	B	309
ICBK	2	Pyrophosphokinase	<i>H. influenzae</i>	A	160	B	160
IQR2	2.1	Quinone reductase type 2	<i>H. sapiens</i>	A	230	B	230
IEV7	2.4	Recombination endonuclease	Bacteriophage T4	A	157	B	157
IEV7	2.4	Restriction endonuclease naei	<i>Nocardia aerocolonigenes</i>	A	295	B	293
IH8X	2	Ribonuclease	<i>H. sapiens</i>	A	125	B	125
II4S	2.2	Ribonuclease III	<i>Aquifex aeolicus</i>	A	147	B	147
IKGN	1.9	Ribonucleotide reductase protein	<i>Corynebacterium ammoniatigenes</i>	A	296	B	296
ITLU	1.6	S-adenosylmethionine decarboxylase	<i>T. maritima</i>	A	117	B	117
IK6Z	2	Secretion chaperone syce	<i>Yersinia pestis</i>	A	120	B	119
IK3S	1.9	Sige	<i>S. enterica</i>	A	106	B	104
IPJQ	2.2	Siroheme synthase	<i>S. typhimurium</i>	A	447	B	454
IHJR	2.5	Site-specific recombinase	<i>E. coli</i>	A	158	C	158
3LYN	1.7	Sperm lysine	<i>Haliotis fulgens</i>	A	122	B	124
2SQC	2	Squalene-hopene Cyclase	<i>Altycyclobacillus acidocaldarius</i>	A	623	B	623
ISCF	2.2	Stem cell factor	<i>H. sapiens</i>	A	116	B	118
IOX8	2.2	Stringent starvation protein B	<i>E. coli</i>	A	105	B	105
PDB		Resolution (Å)	Name	Chain one	Length	Chain two	Length

(continued)

Table 2.8 (continued)

PDB	Resolution (Å)	Name	Source	Chain one	Length	Chain two	Length
1M3E	2.5	Succinyl-coa	<i>S. scrofa</i>	A	459	B	460
1R7A	1.8	Sucrose phosphorylase	<i>Bifidobacterium adolescentis</i>	A	503	B	503
1SOX	1.9	Sulfite oxidase	<i>G. gallus</i>	A	463	B	458
1L5X	2	Survival protein E	<i>Pyrobaculum aerophilum</i>	A	270	B	272
1REG	1.9	T4 rega	Bacteriophage T4	X	122	Y	120
1MKB	2	Thiol ester dehydrase	<i>E. coli</i>	A	171	B	171
1QHI	1.9	Thymidine kinase	<i>Herpes simplex virus</i>	A	304	B	308
1HSJ	2.3	Transcription/sugar-binding protein	<i>E. coli</i>	A	487	B	487
1NY5	2.4	Transcriptional regulator	<i>A. aeolicus</i>	A	384	B	385
1ON2	1.6	Transcriptional regulator	<i>B. subtilis</i>	A	135	B	135
1SMT	2.2	Transcriptional repressor	<i>Synechococcus</i>	A	98	B	101
1TRK	2	Transferase	<i>S. cerevisiae</i>	A	678	B	678
7AAT	1.9	Transferase	<i>G. gallus</i>	A	401	B	401
1KIY	2.4	Trichodiene synthase	<i>Fusarium sporotrichioides</i>	A	354	B	354
1I8T	2.4	Udp-galactopyranose mutase	<i>E. coli</i>	A	367	B	367
1F6D	2.5	Udp-n-acetylglucosamine	<i>E. coli</i>	A	366	B	363
1JP3	1.8	Undecaprenyl pyrophosphate synthase	<i>E. coli</i>	A	210	B	207
1JMV	1.9	Universal stress protein A	<i>H. influenzae</i>	A	140	B	137
1HQO	2.3	URE2 protein	<i>S. cerevisiae</i>	A	221	B	217
9WGA	1.8	Wheat germ agglutinin	<i>Triticum vulgaris</i>	A	170	B	170
1MI3	1.8	Xylose reductase	<i>Candida tenuis</i>	A	319	B	319

Table 2.9 Dataset of homodimers divided into three groups according to their unfolding pathways

PDB ID	Chain	Protein name	Cofactors	Source	ML (aa)
2S (25)					
2cpg	A&B	Transcriptional repressor CopG	–	<i>Streptococcus agalactiae</i>	45
1arr	A&B	Arc repressor	–	Bacteriophage P22	53
1rop	(Sym)	Repressor of protein Rop	–	<i>Escherichia coli</i>	63
5cro	A&C	Cro repressor	–	Bacteriophage lambda	66
1bfm	A&B	Histone B	–	<i>Methanothermus fervidus</i>	69
1a7g	(Sym)	E2 DNA-binding domain	–	HPV strain 16E2	82
1vqb	(Sym)	Gene V protein	–	Bacteriophage f1	87
1b8z	A&B	Histone-like protein HU	–	<i>Thermotoga maritima</i>	90
1ety	A&B	FIS protein	–	<i>E. coli</i>	98
1y7q	A&B	SCAN domain of ZNF 174	–	<i>Homo sapiens</i>	98
1a8g	A&B	HIV-1 protease	–	HIV type 1	99
1siv	A&B	SIV protease	–	SIV	99
1vub	A&B	CcdB	–	<i>E. coli</i>	101
1cmb	A&B	Met repressor	–	<i>E. coli</i>	104
3ssi	(Sym)	Subtilisin inhibitor	–	<i>Streptomyces albogriseolus</i>	108
1wrp	(Sym)	Trp repressor	–	<i>E. coli</i>	108
1bet	(Sym)	β -nerve growth factor	–	<i>Mus musculus</i>	107
1buo	(Sym)	Btb domain from PLZF protein	–	<i>H. sapiens</i>	121
1oh0	A&B	Ketosteroid isomerase	–	<i>Pseudomonas putida</i>	131
2gsr	A&B	Class π glutathione s-transferase	–	<i>Sus scrofa</i>	207
1gsd	A&B	Glutathione transferase A1-1	–	<i>H. sapiens</i>	208
1gta	(Sym)	Glutathione transferase	–	<i>Schistosoma japonica</i>	218
2bqp	A&B	Pea lectin	Mn & Ca ion	Garden pea	234
1hti	A&B	Triosephosphate isomerase	–	<i>H. sapiens</i>	248
1ee1	A&B	Nh(3)-dependent Nad(+) synthetase	–	<i>Bacillus subtilis</i>	271
3SDI (6)					
1mul	(Sym)	Histone-like protein hu- α	–	<i>E. coli</i>	90
1hqo	A&B	Ure2 Protein	–	<i>Saccharomyces cerevisiae</i>	258
1psc	A&B	Parathion hydrolase	Cd ion	<i>Brevundimonas diminuta</i>	329

(continued)

Table 2.9 (continued)

PDB ID	Chain	Protein name	Cofactors	Source	ML (aa)
1cm7	A&B	3-isopropylmalate dehydrogenase	–	<i>E. coli</i>	363
1aoz	A&C	Ascorbate oxidase	Cu ion	Green zucchini	552
1nl3	A&B	SecA	–	<i>Mycobacterium tuberculosis</i>	835
3SMI (10)					
1a43	(Sym)	C-terminal domain of HIV-1 capsid protein	–	HIV type 1	72
1ql1	A&B	Lysine-49 phospholipase A2	–	<i>Bothrops jararacussu</i>	121
1dfx	(Sym)	Desulfoferrodoxin	Fe & Ca ion	<i>Desulfovibrio desulfuricans</i>	125
1yai	B&C	Cu, zn superoxide dismutase	Cu & Zn ion	<i>Photobacterium leiognathi</i>	151
1spd	A&B	Cu, zn superoxide dismutase	Cu & Zn ion	<i>H. sapiens</i>	154
1run	A&B	cAMP receptor protein	–	<i>E. coli</i>	197
1lgs	A&B	Glutathione-s-transferase	–	<i>H. sapiens</i>	209
1tya	(Sym)	Tyrosyl-tRNA synthetase	–	<i>Bacillus stearothermophilus</i>	319
1nd5	A&B	Prostatic acid phosphatase	–	<i>H. sapiens</i>	354
2crk	(Sym)	Creatine kinase	–	<i>Oryctolagus cuniculus</i>	381

ML monomer length, *2S* two-state, *3SDI* three-state with dimeric intermediate, *3SMI* three-state with monomeric intermediate, *SIV* simian immunodeficiency virus, *HIV* human immunodeficiency virus, *HPV* human papillomavirus, *Ccdb* controller of cell division or death B protein, *PLZF* promyelocytic leukemia zinc finger protein, *FIS* factor for inversion stimulation. (sym) indicates that the dimer is generated from a single chain in the PDB by Protein Quaternary Structure Server (PQS) (Henrick and Thornton 1998)

Table 2.10 Data from ASEdb

PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/Mol) ^c	PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/Mol) ^c	PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/Mol) ^c
1a4yA	R5	2.3	1brsA	K27	5.4	1cbwI	P13	–0.1
1a4yA	H8	0.9	1brsA	N58	3.1	1cbwI	K15	2
1a4yA	Q12	0.3	1brsA	R59	5.2	1cbwI	R17	0.5
1a4yA	H13	–0.3	1brsA	E60	–0.2	1cbwI	I19	0.1
1a4yA	R31	0.2	1brsA	E73	2.8	1cbwI	V34	0
1a4yA	R32	0.9	1brsA	R87	5.5	1cbwI	R39	0.2
1a4yA	N68	0.2	1brsA	H102	6	1danA	L39	0
1a4yA	H84	0.2	1brsB	Y29	3.4	1danA	I42	0
1a4yA	W89	0.2	1brsB	D35	4.5	1danA	K62	0
1a4yA	E108	–0.3	1brsB	D39	7.7	1danA	Q64	0.8

(continued)

Table 2.10 (continued)

PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/ Mol) ^c	PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/ Mol) ^c	PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/ Mol) ^c
1a4yA	H114	0.65	1brsB	T42	1.8	1danA	I69	1.9
1a4yB	W261	0.1	1brsB	E76	1.3	1danA	F71	1.2
1a4yB	W263	1.2	1bxiA	C23	0.92	1danA	E77	0
1a4yB	W318	1.5	1bxiA	N24	0.14	1danA	R79	1.2
1a4yB	K320	-0.3	1bxiA	T27	0.73	1danA	Q88	0
1a4yB	E344	0.2	1bxiA	S28	0.17	1danA	V92	0
1a4yB	W375	1	1bxiA	S29	0.96	1danA	N93	0
1a4yB	E401	0.9	1bxiA	E30	1.41	1danA	E94	0
1a4yB	R457	-0.2	1bxiA	L33	3.42	1danA	R271	0
1a4yB	I459	0.7	1bxiA	V34	2.58	1danA	F275	0
1a4yB	Y434	3.3	1bxiA	V37	1.66	1danA	R277	0.51
1a4yB	D435	3.5	1bxiA	T38	0.9	1danA	F278	0
1a4yB	Y437	0.8	1bxiA	E41	2.08	1danA	R304	0.65
1ahwC	Y156	4	1bxiA	S48	0.01	1danA	M306	0.5
1ahwC	T167	0	1bxiA	G49	1.49	1danA	T307	0
1ahwC	T170	1	1bxiA	S50	2.19	1danA	Q308	0
1ahwC	D178	-0.5	1bxiA	D51	5.92	1danA	D309	0.41
1ahwC	T197	1.3	1bxiA	Y55	4.63	1danA	Q312	0
1ahwC	V198	-0.3	1bxiA	P56	1.24	1danA	E325	0
1ahwC	N199	1.1	1cbwI	T11	0.2	1danA	R379	0.51
1danB	K15	-0.4	1danB	E208	0	1dvfB	N55	1.9
1danB	T17	0.1	1dfjI	E202	1	1dvfB	I100	2.7
1danB	N18	0.2	1dfjI	W257	1.3	1dvfB	Y101	4.7
1danB	K20	2.6	1dfjI	W259	2.2	1dvfB	Q103	1.6
1danB	I22	0.7	1dfjI	E283	1.3	1dvfB	R105	4.1
1danB	E24	0.7	1dfjI	S285	0.8	1fc2C	N28	0.6
1danB	Q37	0.55	1dfjI	W314	1	1fc2C	I31	2.2
1danB	K41	0.35	1dfjI	K316	1.3	1fc2C	K35	1.2
1danB	S42	-0.1	1dfjI	E340	1.6	1gc1C	S23	0.29
1danB	D44	0.7	1dfjI	E397	1.3	1gc1C	Q25	0.03
1danB	W45	1.6	1dfjI	Y430	5.9	1gc1C	H27	0.28
1danB	K46	0.25	1dfjI	D431	3.6	1gc1C	K29	0.59
1danB	S47	0.05	1dfjI	Y433	2.6	1gc1C	N32	0.18
1danB	K48	0.4	1dfjI	R453	0.8	1gc1C	Q33	0.1
1danB	F50	0.4	1dfjI	I454	0.3	1gc1C	K35	0.32
1danB	Y51	-0.1	1dvfA	T30	0.9	1gc1C	Q40	-0.41
1danB	D58	2.18	1dvfA	Y32	1.8	1gc1C	S42	0
1danB	D61	0.24	1dvfA	W52	4.2	1gc1C	L44	1.04
1danB	E62	0	1dvfA	D54	4.3	1gc1C	T45	-0.15
1danB	L72	-0.06	1dvfA	N56	1.2	1gc1C	N52	0.7
1danB	F76	1.2	1dvfA	D58	1.6	1gc1C	R59	1.16
1danB	Y78	0.7	1dvfA	E98	4.2	1gc1C	S60	-0.09
1danB	P92	-0.2	1dvfA	R99	1.9	1gc1C	D63	-0.32
1danB	Q110	1.4	1dvfA	D100	2.8	1gc1C	Q64	0.44

(continued)

Table 2.10 (continued)

PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/Mol) ^c	PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/Mol) ^c	PDB ID ^a	Residue ^b	$\Delta\Delta G$ (Kcal/Mol) ^c
1danB	E128	0.1	1dvfA	Y101	4	1gc1C	E85	1.31
1danB	R131	0	1dvfA	H1030	1.7	1jckB	T20	1.4
1danB	L133	0	1dvfA	Y1032	2	1jckB	N23	2.5
1danB	R135	0.55	1dvfA	Y1049	1.7	1jckB	Y26	1.7
1danB	N138	0	1dvfA	Y1050	0.7	1jckB	N60	1.3
1danB	F140	1.5	1dvfA	W1092	0.3	1jckB	Y90	2.5
1danB	S163	0	1dvfA	S1093	1.2	1jckB	V91	2.1
1danB	T203	0.1	1dvfB	H33	1.9	1jckB	G102	0.1
1danB	V207	-0.2	1dvfB	D52	1.7	1jckB	K103	0.4
1jckB	F176	1.9	3hfmA	N93	0.21	3hhrA	I179	0.8
1jckB	Q210	2.5	3hfmA	K96	6.38	3hhrA	R183	0.5
1vfbA	Y32	0.5	3hfmA	K97	5.5	3hhrA	E186	0
1vfbA	W52	1.23	3hfmA	I98	0	3hhrA	P2	-0.05
1vfbA	D54	1.95	3hfmA	S100	0.26	3hhrA	T3	-0.05
1vfbA	R99	0.47	3hfmA	D101	0.94	3hhrA	I4	0.41
1vfbA	D100	3.1	3hfmB	Y58	1.7	3hhrA	R8	0.2
1vfbA	Y101	4	3hfmB	D101	3.75	3hhrA	L9	-0.04
1vfbA	H1030	0.8	3hhrA	H18	-0.5	3hhrA	N12	0.1
1vfbA	Y1032	1.3	3hhrA	H21	0.2	3hhrA	L15	0.15
1vfbA	Y1049	0.8	3hhrA	Q22	-0.2	3hhrA	R16	0.24
1vfbA	Y1050	0.4	3hhrA	F25	-0.4	3hhrA	R19	0.05
1vfbA	T1053	-0.23	3hhrA	D26	-0.2	3hhrA	C182	1.01
1vfbA	W1092	1.71	3hhrA	Q29	-0.6	3hhrB	R43	2.2
1vfbA	S1093	0.11	3hhrA	Y42	0.2	3hhrB	E44	1.8
1vfbB	D18	0.3	3hhrA	L45	1.2	3hhrB	N72	0.2
1vfbB	N19	0.3	3hhrA	Q46	0.1	3hhrB	W76	0.6
1vfbB	Y23	0.4	3hhrA	P48	0.4	3hhrB	T77	-0.25
1vfbB	S24	0.8	3hhrA	S51	0.3	3hhrB	W80	0
1vfbB	K116	0.7	3hhrA	E56	0.4	3hhrB	S98	-0.1
1vfbB	T118	0.8	3hhrA	S62	0.2	3hhrB	S102	-0.2
1vfbB	D119	1	3hhrA	N63	0.3	3hhrB	I103	1.8
1vfbB	V120	0.9	3hhrA	R64	1.6	3hhrB	W104	4.5
1vfbB	Q121	2.9	3hhrA	E65	-0.5	3hhrB	I105	2
1vfbB	I124	1.2	3hhrA	Q68	0.6	3hhrB	C108	0
1vfbB	R125	1.8	3hhrA	Y164	0.3	3hhrB	E120	-0.2
2ptcI	K15	10	3hhrA	R167	0.3	3hhrB	K121	0.1
3hfmA	H15	-0.44	3hhrA	K168	-0.2	3hhrB	C122	0
3hfmA	Y20	4.2	3hhrA	D171	0.8	3hhrB	S124	0.2
3hfmA	R21	0.85	3hhrA	K172	2	3hhrB	D126	1
3hfmA	W63	0.31	3hhrA	E174	-0.9	3hhrB	E127	1
3hfmA	R73	-0.33	3hhrA	T175	2	3hhrB	D164	1.6
3hfmA	L75	0.69	3hhrA	F176	1.9	3hhrB	I165	2.2
3hfmA	T89	0	3hhrA	R178	2.4	3hhrB	Q166	0
3hhrB	R217	0.2	3hhrB	N218	0.3	3hhrB	K167	0
			3hhrB	Q216	0.9	3hhrB	W169	4.5

^aSuffix added to the PDB entry code refers to chain name ^bNumber added to residue symbol is residue position in the chain ^cChange of binding free energy generated in alanine mutagenesis experiment

Table 2.11 Complexes (15) used to generate 296 mutations at the interface

PDB ID	Partner A	Partner B	No. mutations in A	No. mutations in B
1a4y	Angiogenin	Ribonuclease inhibitor	11	12
1ahw	Tissue factor	Immunoglobulin Fab 5G9	7	—
1brs	Barnase	Barstar	7	5
1bxi	Colicin E9 immunity protein	Colicin E9 Dnase domain	17	—
1cbw	Basic pancreatic trypsin inhibitor	Chymotrypsin	7	—
1dan	Blood coagulation factor Viia	Soluble tissue factor	24	34
1dfj	Ribonuclease A	Ribonuclease inhibitor	—	14
1dvf	Fv D1.3	Fv E5.2	16	7
1fc2	Protein A	IgG Fc fragment	3	—
1gc1	Cd4	Envelope protein Gp120	17	—
1jck	Staphylococcal enterotoxin C3	T cell antigen receptor V _β	10	—
1vfb	Monoclonal antibody D1.3	Egg lysozyme	13	11
2ptc	Basic pancreatic trypsin inhibitor	Trypsin	1	—
3hfm	Lysozyme	IgG1 Fab (HyHEL-10)	13	2
3hhr	Human growth hormone	hGH receptor	39	26

kcal mol⁻¹. The dataset was classified into three groups: hot spots ($\Delta\Delta G \geq 1.5$ kcal mol⁻¹), warm residues (0.5–1.5 kcal mol⁻¹) and unimportant residues (<0.5 kcal mol⁻¹), as described by Gao et al. (2004).

2.12 Intronless Genes Dataset

The eukaryotic subdivision files from GenBank are used to create a dataset containing entries that are considered as “intronless” genes according to the “CDS” FEATURES convention (Fig. 2.2). By definition, we consider an entry to be intronless in gene structure if it contains the following description patterns in the corresponding GenBank lines: (1) Contain the word DNA in the LOCUS line at positions 48–53 as per the locus line format; (2) Contain the pattern “CDS” in the FEATURES; (3) The “CDS” line in the FEATURES should contain a continuous span of bases indicated by the number of the first and the last bases in the range separated by two periods (e.g. 23..78). If symbols “<” or “>” are indicated at the end points of the range, the entry is discarded because the range is beyond specified base number in such cases. When operators such as “complement(location)” are used in the “CDS” line, the feature is read as complementary to the location indicated and therefore the complementary strands are read from 5(to 3((Fig. 2.3).

```

FEATURES
    source                Location/Qualifiers
                        1..152254
                        /organism="Chlorokybus atmophyticus"
                        /organelle="plastid:chloroplast"
                        /mol_type="genomic DNA"
                        /strain="SAG 48.80"
                        /db_xref="taxon:3144"
    misc_feature          1..109098
                        /note="LSC; large single copy region"
    gene                 376..1101
                        /gene="ycf27"
    CDS                  376..1101
                        /gene="ycf27"
                        /codon_start=1
                        /transl_table=11
                        /product="regulatory component of sensory transduction
                        system"
                        /protein_id="ABM87971.1"
                        /db_xref="GI:12401235"
                        /translation="MTNTNYNEKIMVVDDEAIVRQIIQTRLISMIGYNVITANDGEEAI
                        KMFFKEQQLVVLDMMPKLDGYSVCQQLRRESQVPIIMLTAEVDADRITGLELGD
                        DYIMKPFSPKELEARIRSVFRSRNRNSYVVKQNKDSSIINIGSLIIDKKRKQVFKDNK
                        RLRLTGMEFSLLELLINRSGESVSRYEILKEVWEYGNDSIDTRVVDVHISRLRSKLE
                        EDPANPDLILTVRGIGYLFQKFD"

```

Fig. 2.2 GenBank FEATURES and CDS annotation (bottom horizontal arrow) for a genomic DNA (top horizontal arrow)

<pre> FEATURES source Location/Qualifiers 1..1497 /organism="Oenothera parviflora" /organelle="plastid:chloroplast" /mol_type="genomic DNA" /cultivar="atrovirens" /db_xref="taxon:482429" /note="plastome type IV" misc_feature 1..1497 /note="large single copy (LSC)" gene 1..1497 /gene="atpB" /locus_tag="OepaCp005" /db_xref="GeneID:5955407" CDS 1..1497 /gene="atpB" /locus_tag="OepaCp005" /codon_start=1 /transl_table=11 /product="ATP synthase beta subunit" /protein_id="YP_001687435.1" /db_xref="GI:169143015" /db_xref="GeneID:5955407" /translation="MRINFTTSGPQVSTLEKKSGRIAGIIGPVLDT NALVVKGRDGGIEHVTCVQQLQNNRVAVANSATDPLTRQREVI GGATLGRIFNVLGEVDLGPVDTKTTSPINRSAPAFIQLDKLSIFE PYRRGGKIGLFGGAGVGKTVLIMELININIAKGGVSVFGVGGERTE SGVINEQNIASKVALYVGQNNPPGARMRVGLTALTRAEYFRDNNKQ RFVQAGSEVSAALLGRNPSAVGVQPTLSTMGSLQERITSTKAGSITSI TDPAPATTFANLDAITVLSRLAAGKGYAVDPLDSTSTMLQPRIVGD ETLQRYKELQDIISILGDELSEEDRLTVARARKIERFLSQPFVFAEV LAETIROPKLLILSGELGDPGAFLVTVTIDEATAKANLEHESLKK 1494..>1497 gene 1494..>1497 /gene="atpE" /locus_tag="OepaCp006" /db_xref="GeneID:5955467" CDS 1494..>1497 /gene="atpE" /locus_tag="OepaCp006" /codon_start=1 /transl_table=11 /product="ATP synthase epsilon subunit" /protein_id="YP_001687436.1" /db_xref="GI:169143016" /db_xref="GeneID:5955467" /translation="MTNLCLVLTTPHIVVDSEVKEIILSTNSGOIGVLPNHPAIATAV DIGILIRIQNGQWLTALMGGFARIGNNEITLVNDAEKQSDIDPQEAQETLGLAEAN FRKAEGRQTTIEANLALRRARTVEAINVIS" </pre>		<pre> FEATURES source Location/Qualifiers 1..1428 /organism="Oenothera parviflora" /organelle="plastid:chloroplast" /mol_type="genomic DNA" /cultivar="atrovirens" /db_xref="taxon:482429" /note="plastome type IV" misc_feature 1..1428 /note="large single copy (LSC)" gene 1..1428 /gene="cbcl" /locus_tag="OepaCp004" /db_xref="GeneID:5955425" CDS 1..1428 /gene="cbcl" /locus_tag="OepaCp004" /codon_start=1 /transl_table=11 /product="tribulose-1,5-bisphosphate c large subunit" /protein_id="YP_001687434.1" /db_xref="GI:169143014" /db_xref="GeneID:5955425" /translation="RSFQTETKASVGFKAQGVYKLT QGVPPPEEACAAVAEESSTQTTVTFTDGLTSLDRYK YPLDLFEQGVTHNFTIVQNVFGFKALRALREDLR KLNKYGRPLLGCTIEPKLGLSKNYGRAVYECLELGG LFCAEIYKSQAETGEIKGHYLNATAGTCCEMMKRAI NTSLANHYCRDNGLLHHRANHAVIDRQNNHGHFRV LEGEDITLGFVLLDRDIEKDRSRGIYFTQDVSGL IFGDDSVLFGGGTGLHPWQNAQAVANRVALEACVQ </pre>
--	--	--

Fig. 2.3 CDS annotation for direct, complement and partial intronless genes

Table 2.12 Different CDS (coding sequence) patterns used for SEG annotation in GenBank FEATURES format is shown

Entries	Nature	CDS patterns for SEG	No. of entries	Total entries
Complete	Direct	a..b	1,903	3,750
	complement	complement(a..b)	1,847	
Partial	Direct	<a..>b	140	184
	complement	complement(<a..>b)	44	

2.13 Human Single Exon Gene (SEG) Dataset

Human SEG sequences were obtained from the Genome SEGE database (Sakharkar and Kanguane 2004) created using a procedure described above in Sect. 2.12. This procedure utilized CDS annotation in the FEATURES (GenBank formatted record) for the identification and extraction of SEG sequences from the human genome. The CDS annotation in the FEATURES contains several patterns (complete [direct or complimentary] or partial [direct or complimentary]) for representing SEG and these patterns are summarized in Table 2.12. Thus, we obtained 3,750 SEG nucleotide sequences from the human genome. The human genome file does not contain protein translations. A protein translation file called “protein.fa” (file containing all protein sequences in the human genome) containing protein sequences was downloaded. The “protein.fa” file contains protein translation for 3,656 SEG sequences. These SEG protein sequences formed a dataset for human SEG.

2.14 Intron-Containing Genes Dataset

The data is obtained from GenBank for the dataset. The invertebrate, mammalian, plant, primate, rodent and vertebrate subdivisions, which represent all the eukaryotic gene entries, are considered. The information on sequence, length, position and intron phase at intron–exon junction was obtained by parsing the CDS features in the GenBank records (Fig. 2.4). The protein sequence from the GenBank records, protein ID, product information and gene name are extracted from the translation, protein ID and product/gene name qualifiers, respectively, present in the CDS field of the FEATURES. If the 30 and 50 ends of the flanking exons for an intron are available in different GenBank records – as in the segmented genes – then the entries are marked as fusion entries and the positions of introns are marked by an “@” sign.

Partial sequences are identified based on the “<” or “>” symbol in the CDS field. Since there is an error rate in the prediction of the intron positions by computer programs, we decided to create subsets, one containing all entries where the introns were predicted and another one containing entries where the introns were characterized by experiments. Entries containing predicted introns were identified by searching

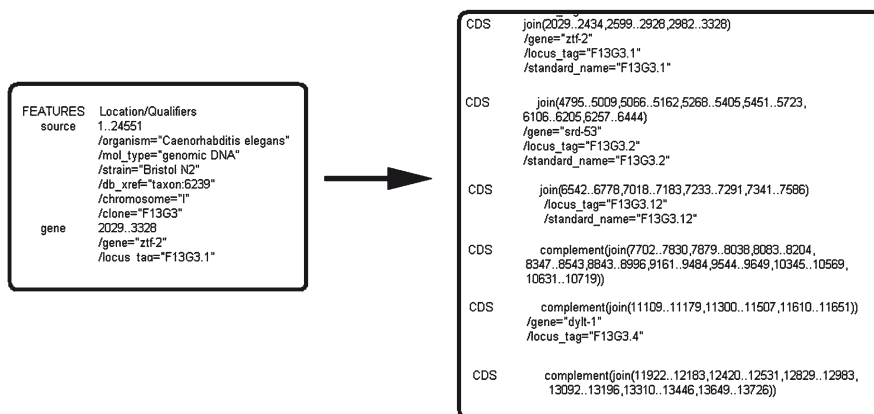


Fig. 2.4 Different CDS representations for intron-containing multiple exon genes in eukaryotes are illustrated

for the words “cosmid,” “BAC,” “PAC” or chromosome. The dataset is also divided into two other independent subsets containing the entries corresponding to organelle and nuclear genes by searching for words “mitochondrial,” “chloroplast” or “plastid” in the “ORGANISM” line of GenBank entry.

2.15 Fusion Protein Dataset

Gene fusion has been described as an important evolutionary phenomenon (Fig. 2.5). This report focuses on identifying, analyzing, and tabulating human fusion proteins of prokaryotic origin. These fusion proteins are found to mimic operons, simulate protein–protein interfaces in prokaryotes, exhibiting multiple functions and alternative splicing in humans. The 26,673 nonhomologous human protein set (created by removing similar sequences at 40% sequence similarity criteria) are searched against the 102,135 nonhomologous bacterial protein set (created by removing similar sequences at 40% sequence similarity criteria) using BLASTP (protein sequence search software – <http://blast.ncbi.nlm.nih.gov/Blast.cgi>) at an E value (expect value) cutoff of 10^{-10} . This experiment identified human fusion proteins consisting of two or more fusion partners of prokaryotic origin. The list is available elsewhere (Yiting et al. 2004).

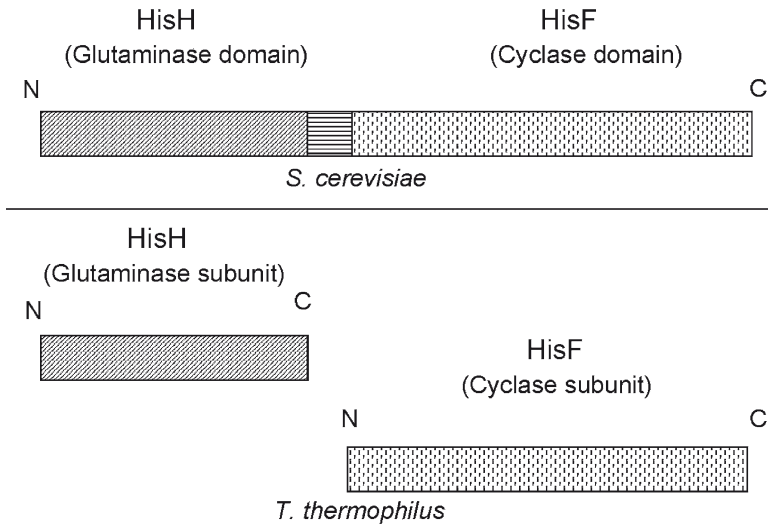


Fig. 2.5 Fusion protein scenario for imidazole glycerol phosphate synthetase (IGPS) in yeast and bacteria

2.16 Exercises

1. Examine PDB and create a dataset for protein–DNA and protein–RNA complexes.
2. Give a statistics of HLA alleles in IMGT/HLA database.
3. Give a quantitative account for data statistics in major genetic databases.
4. Create an updated dataset of MHC–peptide complexes from PDB.
5. How many monomers and multimers are solved and made available in PDB? Give an account.
6. Relate sequence data in GenBank and structure data in PDB with a statistical account.
7. What are the salient aspects of the FEATURES in GenBank?
8. Illustrate the CDS annotations for intron-containing and intronless eukaryotic genes in GenBank.
9. What are the different methods used to represent partial CDS in GenBank?
10. Give an account of binding data for biologically important macromolecules.

References

- Alexander J, Del Guercio MF, Fikes JD, et al. Recognition of a novel naturally processed, A2 restricted, HCV-NS4 epitope triggers IFN-gamma release in absence of detectable cytopathicity. *Hum Immunol.* 1998;12:776–782.
- Chang KM, Gruener NH, Southwood S, et al. Identification of HLA-A3 and HLA-B7-restricted CTL response to hepatitis C virus in patients with acute and chronic hepatitis C. *J Immunol.* 1999;162:1156–1164.
- Chen W, Khilko S, Fecondo J, et al. Determinant selection of major histocompatibility complex class I-restricted antigenic peptides is explained by class I-peptide affinity and is strongly influenced by nondominant anchor residues. *J Exp Med.* 1994;180:1471–1483.
- Den Haan JM, Meadows LM, Wang W, et al. The minor histocompatibility antigen HA-1: A diallelic gene with a single amino acid polymorphism. *Science.* 1998;279:1054–1057.
- Gao Y, Wang R, Lai L. Structure-based method for analyzing protein-protein interfaces. *J Mol Model.* 2004;10:44–54.
- Gianfrani C, Oseroff C, Sidney J, et al. Human memory CTL response specific for influenza A virus is broad and multispecific. *Hum Immunol.* 2000;61:438–452.
- Henrick K, Thornton JM. PQS: a protein quaternary structure file server. *Trends Biochem Sci.* 1998;23:358–361.
- Kawashima I, Hudson SJ, Tsai V, et al. Multi-epitope approach for immunotherapy for cancer: identification of several CTL epitopes from various tumor-associated antigens expressed on solid epithelial tumors. *Hum Immunol.* 1998;59:1–14.
- Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A.* 1996;93:13–20.
- Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J Mol Graph.* 1995;13:323–330.
- Lauvau G, Kakimi K, Niedermann G, et al. Human transporters associated with antigen processing (TAPs) select epitope precursor peptides for processing in the endoplasmic reticulum and presentation to T cells. *J Exp Med.* 1999;190:1227–1240.
- Lee B, Richard FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol.* 1971;55:379–400.
- Livingston BD, Crimi C, Fikes J, et al. Immunization with the HBV core 18–27 epitope elicits CTL responses in humans expressing different HLA-A2 supertype molecules. *Hum Immunol.* 1999;60:1013–1017.
- McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol.* 1994;238:777–793.
- Nukaya I, Yasumoto M, Iwasaki T, et al. Identification of HLA-A24 epitope peptides of carcino-embryonic antigen which induce tumor-reactive cytotoxic T lymphocyte. *Int J Cancer.* 1999;80:92–97.
- Rechenmann F. From data to knowledge. *Bioinformatics.* 2000;16:411.
- Sakharkar MK, Kanguane P. Genome SEGE: A database for ‘intronless’ genes in eukaryotic genomes. *BMC Bioinformatics.* 2004;5:67.
- Service RF. Structural genomics offers high-speed look at proteins. *Science.* 2000;287:1954–1956.
- Sette A, Sidney J, del Guercio MF, et al. Peptide binding to the most frequent HLA-A class I alleles measured by quantitativemolecular binding assays. *Mol Immunol.* 1994;31:813–822.
- Yiting Y, Chaturvedi I, Liew KM, et al. Can ends justify the means? Digging deep for human fusion genes of prokaryotic origin. *Front Biosci.* 2004;9:2964–2971.



<http://www.springer.com/978-1-4419-0519-2>

Bioinformation Discovery

Data to Knowledge in Biology

Kangueane, P.

2009, XXV, 166 p.,

ISBN: 978-1-4419-0519-2