Russian Chemical Reviews **74** (1) 61 – 76 (2005) © 2005 Russian Academy of Sciences and Turpion Ltd

DOI 10.1070/RC2005v074n01ABEH001159

Catalytic sites of hydrolases: structures and catalytic cycles

S D Varfolomeev, I A Gariev, I V Uporov

Contents

I.	Introduction	61
II.	Databases of amino acid sequences and three-dimensional structures of enzymes	61
III.	Methods for identification of groups constituting the catalytic site	63
IV.	Catalytically important groups and structure-forming amino acids in the active sites of enzymes	65
V.	Classification of hydrolases according to the types of catalytic sites and catalytic mechanisms	66
VI.	Glycosyl hydrolases, cellulases	69
VII.	Serine hydrolases. Automatic search for a catalytic triad	70
VIII.	Mechanism of activation of substrates by active sites of hydrolases	71
IX.	Glycine, cysteine and proline as structure-forming amino acids	72
X.	The rates of conformational changes and proton transfer as the limiting steps	73

Abstract. The applications of various bioinformatics methods and databases to the analysis and classification of proteins and search for active sites of enzymes are considered. The published data on the structures of active sites and mechanisms of action of hydrolases are analysed. It is emphasised that all of the known hydrolases can be divided into four groups according to the types of catalytic site which differ in the mechanisms of water activation. The mechanisms of water activation by catalytic groups in the active sites of hydrolases are discussed and the schemes of catalytic cycles characteristic of classical types of active sites of these enzymes are considered. Using glycosyl hydrolases and serine hydrolases as examples, it is demonstrated that different proteins possess common catalytic sites and mechanisms of action. The role of aspartic acid and histidine in the activation of hydrolase substrates and the role of structure-forming acids (glycine, proline, cysteine) in the structural organisation of the catalytic sites are considered. The bibliography includes 195 references.

I. Introduction

Considerable progress achieved in the study of enzymes in recent decades is caused by the profound fundamental interest in enzymes as highly efficient catalysts of chemical reactions, on the one hand, and by continuously expanding areas of their applications aimed at the solution of numerous practical tasks, on the other hand.[1 – 13]

The large amount of information on the structures of enzymes has been accumulated over recent years, considerable progress has been observed in the development of concepts of the principles of formation of active sites of enzymes and their mechanisms of action. Tremendous amounts of information concerning amino acid sequences and spatial structures of protein molecules have been amassed in numerous computer databases. This information demands analysis, systematisation and interpretation from the standpoint of the chemistry of catalytic processes. This review is devoted to applications of bioinformatics methods and databases in the analysis of enzymes and search for their active sites.

Hydrolases, *i.e.*, enzymes catalysing hydrolysis of various compounds, were chosen as the subjects of investigation. This choice was determined by a number of circumstances. Hydrolases constitute about one third of all currently known enzymes (approximately 1100 out of 3700 in the Enzyme Classification). The functional groups in the active sites of many hydrolases that effect their catalytic cycles have been determined. The mechanisms of action of all hydrolases include activation of water by a limited array of functional groups. The fact that many structural elements in the active sites of hydrolases are present in enzymes pertaining to other classes is of no less importance for enzymatic catalysis at large.

II. Databases of amino acid sequences and three-dimensional structures of enzymes

A considerable body of information on the primary and spatial structures of biopolymers has been accumulated to date. Information on amino acid sequences of enzymes is widely used for their classification, *i.e.*, belonging to one or another family,† determination of conserved amino acids and their positions within the families, studies of translated sequences of proteins derived from decoded genomes, *etc*. Spatial structures of enzymes are used for the studies of enzyme-catalysed chemical transformations, rationalisation of substrate specificities, site-directed mutagenesis of enzymes aimed at conferring specific functions on them, *etc*.

Databases containing amino acid sequences or spatial structures of proteins are referred to as primary. Secondary databases comprise elements of the primary databases classified according to their specific features, *e.g.*, databases on alignments of amino acid sequences of proteins or hierarchical databases of protein structures.

S D Varfolomeev, I A Gariev, I V Uporov Department of Chemistry, M V Lomonosov Moscow State University, Leninskie Gory, 119992 Moscow, Russian Federation. Fax (7-095) 939 54 17, tel. (7-095) 939 35 89, e-mail: sdvarf@enzyme.chem.msu.ru (S D Varfolomeev), tel. (7-095) 939 34 07, e-mail: gariev@hotmail.com (I A Gariev), e-mail: iuporov@rabbit.chem.msu.ru (I V Uporov)

Received 23 July 2004
Uspekhi Khimii **74** (1) 67 – 83 (2005); translated by R L Birnova

† The terminology used by different authors may differ.

The Swiss-Prot/TrEMBL (EMBL stands for 'European Molecular Biology Laboratory' database),[14] which further developed into the knowledgebase UniProt,[15] is one of the most popular primary databases of protein sequences. The non-commercial database Swiss-Prot is supported by several research institutes in Europe (*e.g.*, the Swiss and European Institutes for Bioinformatics) and the USA and is accessible through ExPASy (Expert Protein Analysis System), the proteomic server of the Swiss Institute for Bioinformatics (http://www.expasy.org). Information deposited therein is available through servers in many other countries, which ensures rapid responses to the users' queries.

By the time of writing of this paper (June 2004), the latest version of the Swiss-Prot 43.5 database contained information about 153 017 annotated amino acid sequences with the overall length of 56 301 572 amino acids retrieved from 116 894 publications. It is noteworthy that this database contains only annotated sequences, *i.e.*, sequences of the proteins (including enzymes) with well-established functions. The database TrEMBL, which comprises translated protein sequences, is a supplement to the Swiss-Prot database. The translation is carried out automatically from the nucleotide sequences deposited with the EMBL Nucleotide Sequence Database.[16] For the protein sequences deposited in the database TrEMBL, only the sources are known. The conclusion about their putative functions is made on the basis of their similarities to the sequences of the studied (annotated) proteins. The latest version of the database TrEMBL contained information about 1 062 416 sequences with the overall length of 332 514 237 amino acids. Special mention should be made of a nearly tenfold difference between the number of protein sequences in the databases Swiss-Prot (annotated) and TrEMBL (translated). As a result of their investigation, the proteins from the database TrEMBL pass into the database Swiss-Prot, where they supplement the arrays of sequences of the annotated proteins. The number of sequences in the database Swiss-Prot/TrEMBL increases nearly exponentially with time.

The data deposited with the database Swiss-Prot/TrEMBL encompass several fields; their content can be illustrated by the following example. The amino acid sequence of formate dehydrogenase from *Pseudomonas sp*. strain 101 has the primary accession number P33160 (a unique identifier of this sequence in the Swiss-Prot) and the entry name FDH_PSESR (the second unique identifier of this particular sequence having a definite semantic meaning): FDH is abbreviated from Formate DeHydrogenase, while the abbreviation PSESR indicates the source, *viz*., *Pseudomonas sp*. Using the accession number or the entry name, one can obtain available information about the amino acid sequence.‡ The accession number and the entry name (if it exists) are followed by chronological information (database entry date, latest modification date, annotation modification date, *etc*.); name(s) of the protein [official, synonyms, Enzyme Classification (EC) number]; origin and taxonomic data; references to the literature sources of the sequence or its corrections; comments on enzyme-catalysed reactions and data on the quaternary structure and relatedness to a protein family. Other information includes cross-references to the corresponding entries of other databases. For formate dehydrogenase used as an example, there are references to the entry in the protein structure database PDB (Protein Data Bank, 2NAC identifier) and two entries in the Protein families database (Pfam).[17] It is of note that the number of cross-references varies for different sequences and depends on the level of investigation of a protein. The cross-references are succeeded by the field 'Features', which provides information about the domain structure and secondary structure elements based on the spatial structure of a protein (if any) and, finally, the amino acid sequence itself.

For the overwhelming majority of protein sequences deposited in the database Swiss-Prot, the set of fields is not that big, since the spatial structures are known only for a limited number of proteins including well-studied ones. For protein sequences deposited with the database TrEMBL, the set of fields is even smaller: it is only the sequence source (gene), references and the sequence itself that are reliable. It is of note that the access to the database Swiss-Prot/TrEMBL permits the use of different kinds of search and the user's interface is well thought-out, which significantly facilitates the work.

The Protein Data Bank (PDB)[18] (http://www.rcsb.org) represents a primary database of spatial structures of biopolymers. This data bank is supported by the Research Collaboratory for Structural Bioinformatics which comprises three research centres in the USA. By June 2004, this data bank comprised information about 25 882 biopolymer structures (including 23 429 peptide and viral structures) established by X-ray diffraction analysis and multidimensional NMR.§ It is of note that the number of structures deposited with PDB increases exponentially and so does the volume of data banks of the primary sequences. However, in contrast to amino acid sequence data banks, the structure data bank includes different structures of the same biopolymer obtained under different conditions (*e.g.*, pH) or as a complex with different molecules (*e.g.*, with inhibitors), *etc*. Some protein structures look very similar. If such 'similar' structures are omitted, the remaining $\sim$ 5 000 structures will differ, in one degree or another, in the fold of their polypeptide chains, the increase in the number of new folding types being far from exponential. It is noteworthy that the difference in the growth rates of the bank of primary sequences and the bank of spatial protein structures makes the simulation of protein structures based on their amino acid sequences a topical task. The development of methods for simulation is progressing rapidly.[19]

Every structure deposited with the PDB is presented as a text file, which includes the coordinates of all of its heavy atoms (these coordinates are read by molecular graphics programs or are used as the starting information for calculations). Some proteins possess spatial elements (groups of atoms) manifesting high lability (these include disordered structures situated on the surface of protein globules or several amino acid residues nearest to the N-/C-ends). The impossibility of determining their coordinates leads to the appearance of gaps in the polypeptide chains or a lack of structural data on the side chains of amino acid residues.

In addition to the atom coordinates, PDB files contain references to the papers with the protein structures, experimental procedures and subsequent processing of the results obtained, amino acid sequences and secondary structures of proteins. High accuracy of determination of spatial structures of enzymes makes quantum-chemical methods applicable to simulation of enzymatic catalytic processes.[20]

The interface of the PDB database allows text searching in different fields. In our opinion, the database PDBsum (http://www.biochem.ucl.ac.uk/bsm/pdbsum/) developed and supported by researchers of the European Institute for Bioinformatics is the most convenient tool for object searching.[21–23] This database is secondary according to our classification and contains references to PDB files ordered according to specific features. For example, it allows a rapid search for protein structures on the basis of their EC numbers, determination of all structures containing a specific ligand or a prosthetic group (*e.g.*, heme or NAD) and retrieval of a list of biomolecules isolated from any particular organism. In addition to a large number of cross-references, this database provides a graphical presentation of the secondary structure of a protein with the indication of its conserved amino acids.

‡ It is of note that the primary accession number and entry name for each amino acid sequence deposited with the database TrEMBL coincide, since their functions are unknown and the conferment of names reflecting their function or classification is unjustified.

§ The term 'biopolymer (particularly, protein) structure' stands for a set of Cartesian coordinates of heavy atoms (of all atoms with the exception of hydrogen atoms) constituting the given biopolymer.

Two databases, *viz*., CATH (http://www.biochem.ucl.ac.uk/bsm/cath/),[24–25] and SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/),[26–28] comprise a hierarchical classification of protein structures. The structural domains of the proteins deposited with these databases are clustered according to the fold of the polypeptide chain. The hierarchical nature of these databases is manifested in the stepwise introduction of similarity levels of protein structures. Thus the structures in the database CATH are first combined into homologous superfamilies (Homology), the families are united into topologies (Topology), which form the topology architectures (Architecture); the latter, in turn, form classes (Class). A knowledge of the index of a structure in the database classification allows a rapid search for structural 'neighbours' and determination of the qualitative degree of similarity between different protein structures.

The database HSSP (Homology-derived Secondary Structures of Proteins)[29] available at the address http://www.sander.ebi.ac.uk/hssp/ is a database linking structural information to sequence information. This database comprises pairwise alignment of the amino acid sequence of a protein deposited with the data bank PDB and homologous sequences with the degree of identity of 30% and higher from the data bank Swiss-Prot/TrEMBL. This information can be used for the construction of spatial structure models of proteins using homologous simulation.[19] The database HSSP also contains information about the conservation of the constituent amino acids in the amino acid sequences.

The aforementioned database Pfam is the only secondary database of protein sequences that deserves special consideration.[17] This database comprises multiple alignments of protein sequences from the database Swiss-Prot/TrEMBL, united into families according to the similarity of their amino acid sequences. The protein families deposited with the version 14.0 of this database (June 2004) cover 95% of all known protein sequences. The database Pfam is available at the address http://www.sanger.ac.uk/Software/Pfam/. Multiple alignments of protein sequences have made it possible to single out domains manifesting the highest degree of similarity. The amino acid sequence of a real protein may contain several domains. It is therefore important to discriminate between the structural domains of the proteins deposited with the databases CATH and SCOP and the amino acid sequence domains. The former are derived from similar folding of the peptide chains of proteins with the known spatial structures, while the latter were established by multiple alignments of an array of protein sequences. Very often, the boundaries of structural and amino acid sequence domains coincide.

Multiple alignments in the database Pfam were built by the program HMMER (http://hmmer.wustl.edu) based on the Hidden Markovian Model.[29] The alignment of sequences by the HMM method is characterised by an extremely high sensitivity in the determination of homologies between evolutionary distant proteins. This method is based on the concept according to which every position in a sequence is determined by a random Markovian process, which results in the generation of an amino acid. In this approach, conserved positions generate only one type of amino acid, while less conserved positions generate two or three such types. In mathematical terms, this fact can be described by the introduction of probabilities of generation of amino acids in all positions. Moreover, it is necessary to introduce probabilities of a transition from one position to another, gap probabilities and so on. The set of such probabilities and the HMM architecture constitute the main principle of the HMM method for the given protein family. Some efficient methods for the determination of a set of probabilities by sequential iteration have been devised. The main advantage of the HMM method over traditional alignment methods is that the penalty of the sequence gap is not an external parameter, but rather is established in a self-consistent mode in the course of HMM construction and varies on going from one position to another. The HMM method is successfully used for predicting structures and hypothetical functions of protein sequences in the genome annotation.[30] In addition to HMM parameters, the database Pfam contains the multiple alignment itself and the corresponding dendrogram (a phylogenetic tree), *i.e*., a schematic presentation of the spatial structure of the given domain (if a family comprises at least one domain with a known spatial structure) and a histogram of probabilities of amino acid generation in each position of the protein domain (the higher the column height, the higher the conservation of the given position). This information is extremely important in designing site-directed mutagenesis of enzymes, since it allows rapid localisation of the most significant amino acid residues for the tested sequence.

The list of the aforementioned databases of protein sequences and structures is far from being complete. We have mentioned and discussed only some of these databases which, in our opinion, comprise important information about amino acid sequences and spatial structures of enzymes. This information is displayed in a convenient form and is extremely useful in the different steps of an investigation.

III. Methods for identification of groups constituting the catalytic site

Information about the involvement of one or another group in the catalytic site and in the catalytic process at large can be derived from different (often complementary) sources. An essential role in the development of concepts on the structure of active sites is played by the methods of chemical modification of amino acid residues, especially if a modifier can be incorporated into a substrate or an inhibitor to be further directed to the active site of the enzyme.[4,11] The use of site-directed mutagenesis with replacement of one or several amino acids in the polypeptide chain of the enzyme enables reliable identification of participants in the catalytic process. This methodological approach is very popular.[31–36] The design and progress in the development of methods for the studies of spatial structures of proteins has provided this branch of enzymology with abundant and highly reliable structural information.[37–40] At the same time, this information is insufficient for identification of groups constituting the active site. It is necessary to isolate a fragment directly involved in the catalytic process. The approaches to the solution of this task are discussed below.

Owing to the decoding of DNA sequences of individual genes and entire genomes, the number of proteins with established amino acid sequences exceeds 1 000 000. However, this diversity is rather apparent, since different organisms often contain similar (homologous) proteins derived from the same evolutionary progenitor. Their amino acid sequences are characterised by an extremely high degree of identity. In the case of enzymes, they possess similar catalytic sites, mechanisms of action, cofactors and inhibitors. Therefore, the bioinformatics methods are often employed to classify enzymes based on the analysis of their primary structures and to unite them into homologous protein families. It should be noted that methods of comparison (alignment of amino acid sequences and detection of the degree of their identity) are fairly well-developed and automated; therefore, this approach allows operation with large amounts of information in an automatic mode. The two main databases using this approach for classification of hydrolases are MEROPS[41–50] and CAZY.[51–59]

The database MEROPS[41–50] (http://merops.sanger.ac.uk/) comprises information about peptidases and their inhibitors. Two enzymes are united into a family, if the degree of identity of their amino acid sequences exceeds 30%. In turn, the families are united into clans, *viz*., groups of families where protein homology cannot be established by analysis of their amino acid sequences, although there is some indirect evidence (*e.g*., the fold of the polypeptide chains and the sequence order of the active site residues in amino acid sequences), which suggests a relationship between these proteins. The clans are united into classes according to the type of their catalytic mechanisms. For example, serine, thiol, aspartyl

and metal-dependent proteases are known; the serine proteases are divided into 10 clans which in turn are divided into 25 families that comprise altogether 6 886 proteins.

The database CAZY [51–59] (http://afmb.cnrs-mrs.fr/CAZY/) contains information about glycosidases, glycosyl transferases and other enzymes catalysing the formation and cleavage of the glycosidic bond. Based on the results of analysis of identity of their amino acid sequences, all glycosidases are divided into 14 clans which unite 96 families.

An analysis of identities of amino acid sequences was also employed in the construction of the database Catalytic Site Atlas [60, 61] (http://www.ebi.ac.uk/thornton-srv/databases/CSA/), although inclusion of enzymes into homologous protein families was not their main goal. The base designers have chosen about 300 most well-studied enzymes with established three-dimensional structures. The data on the amino acid residues constituting the active sites and a list of proteins with amino acid sequences with high degree of identity are given separately for each enzyme.¶

The bioinformatics methods of analysis of amino acid sequences and determination of protein functions considered above have become very popular owing to abundant original structures obtained by translation of their nucleotide sequences. However, the functional potentials of proteins are determined primarily by their three-dimensional (rather than primary) structures. Although high degree of identity of amino acid sequences underlies the similarity of spatial structures of active sites of enzymes, in the general case the converse is wrong. Thus the catalytic triad Ser-His-Asp is characteristic of both trypsin and subtilisin. These proteins possess non-homologous amino acid sequences, they differ in the order of amino acid residues within the active site and the fold of the protein globule. Mammalian trypsin and bacterial subtilisin provide examples of convergent evolution of enzymes where enzymes derived from different proteins acquire similar (in all probability, optimum) structures of catalytic sites in the course of evolution. However, a comparison of primary amino acid sequences of such proteins does not allow one to establish a similarity of their catalytic sites.

The methods for analysis of three-dimensional (3D) structures of proteins allow one to cope, in principle, with evolutionally non-related proteins; however, they are much less developed than the methods for the analysis of amino acid sequences and are still at the stage of development rather than at the wide-scale application stage. First, this is due to the complexity of the systems under study and, second, to the lower level of the original data (since proteins with established three-dimensional structures are far less numerous than proteins with established primary structures). It can be assumed that in future the methods of analysis of three-dimensional structures of proteins will play a more significant role, since the number of structures in the database PDB is steadily increasing and so is the accuracy of prediction of the 3D-structures of proteins based on their amino acid sequences. Some methods aimed at a search for functional sites on the basis of structural data have been reviewed.[63]

A convenient method for classification of active sites of enzymes and substrate-binding sites (ligands or cofactors) according to the three-dimensional structure of proteins has been developed.[64] Considering that the active sites of 70% of enzymes are localised in the largest cavity of the protein globule,[65] one can restrict oneself to the analysis of the amino acid sequences of the protein that form the cavity surface. Such residues can be found automatically. The search can culminate in the physicochemical 'image' of the cavity, *viz*., spatial arrangement of charges, the groups serving as potential donors or acceptors of the hydrogen bond as well as hydrophobic and hydrophilic sites on the protein surface. The presentation of the cavity as a graph with the atoms as vertices and the distances between them as edges allows a comparison of two proteins irrespective of their amino acid sequences as well as a search for the greatest common elements using an algorithm for the search of the largest common subgraph. This approach was used for identification of substrate-binding sites in subtilisin, trypsin and chorismate isomerase isolated from different sources with the degree of identity of the amino acid residues of $< 20\%$, as well as NADH-binding sites in several non-related oxidoreductases.

Although this method is applicable to the study of substrate-binding sites, it does not seem to be suitable for the search for the catalytic sites themselves. For instance, it is only the serine residue in serine hydrolases with the catalytic triad Ser-His-Asp that is localised on the surface of the protein globule, while histidine and aspartic acid residues are embedded inside the protein.

The so-called method of geometrical invariants was suggested for classification of proteins according to the presence of clusters built of several amino acid residues with similar spatial distributions.[66] All possible combinations of several amino acid residues are sorted out for each protein; the geometrical invariants, *i.e*., the parameters dependent on the coordinates of the appropriate atoms, but independent of translations and frame rotations are calculated. These include the distance between two atoms, the volume of a tetrahedron formed by four atoms, *etc*. For each group of residues, there is a single point in the *n*-meric space (*n* is the selected number of invariants) and a set of such points for each protein. By applying this procedure to all proteins, one can obtain several domains in the *n*-meric space where the points corresponding to different proteins are concentrated. All these proteins are characterised by a similar arrangement of amino acid residues within the protein structure. This method has made it possible to determine the active site residues for many enzymes including cysteine proteases, xylanases and β-lactamases. The advantage of this method is that it permits an automated search for common structural elements for different proteins and does not require preliminary elucidation of common elements of the tested proteins. Besides, it enables the detection of common elements even in the sites where their occurrence was thought to be hardly probable.

In yet another approach, which is based on expert knowledge and analysis of the published data, a desired spatial arrangement of amino acid residues or individual atoms,† which corresponds, *e.g*., to the active site of the enzyme, is preset, and the search routine finds the corresponding proteins in the structure database.

The graph-theoretic approach based on the use of templates for the classification of proteins has been suggested.[67, 68] Each amino acid residue is represented by a vector between two atoms depending on the type of the amino acid residue. The reciprocal spatial arrangement of two vectors is determined by the distances between their initial, final and central points. A template represents an array of the vectors and a matrix of triples of distances between them plotted as a graph; each graph vertex corresponds to an amino acid residue and each edge is a set of distances between the vectors of the residues. One graph isomorphism search algorithm is used as a template in a search for protein structure coincidence. This method has been realised in the program ASSAM, which was successfully used for a correct search of the catalytic triad Ser-His-Asp, the zinc-binding site of thermolysin (and related enzymes) and a great number of other functional sites of proteins.

The transition from the presentation of an amino acid residue as a vector (which imparts an additional degree of axial symmetry to the amino acid residue) is an obvious route to further perfection of the method. This approach was realised in the programs TESS and JESS.[69–71] The coordinates of the atoms from the primary database PDB of amino acid residues with established functional

¶ The search for the appropriate proteins was carried out using the BLAST method (Basic Local Alignment Search Tool).[62]

† The set of limitations for the types of amino acid residues or atoms and their reciprocal spatial distributions is usually referred to as a template, although the terminology used by different authors may differ.

significance were used as templates. This approach resulted in the construction of the database PROCAT[70] (http://www.biochem.ucl.ac.uk/bsm/PROCAT/), which comprises the data on the active sites in protein structures.

The aforementioned methods make use of the data only on three-dimensional structures of proteins. They seemed opposed to the methods based only on the primary structures data. In actual research practice, both approaches are used to determine the functional significance of amino acid residues.

The Evolutionary Trace method[72] consists of the search (usually automated, employing the BLAST method) of a set of homologous proteins, their alignment and construction of a dendrogram (a tree) from the alignment data obtained. Each terminal node in the tree corresponds to a real protein and each intermediate node corresponds to a certain hypothetical protein called 'evolutionary precursor'. In the sequence of the latter, only that part of the amino acid sequence is known which is common for all the proteins of the daughter nodes. Therefore, the amino acid residues of the starting protein can be ranged from the most conserved, identical in all homologous proteins of the set, to the most variable ones, which are different in the original protein and its closest relative. The depth of the node in the dendrogram where the divergence took place is a quantitative measure of amino acid residue conservation. Having removed the residues below a certain degree of conservation, one can isolate clusters of residues which, according to the structural data, are in contact with one another. A correlation exists between the conserved clusters and the functional sites of the protein. The method allows correct determination of 70% – 90% of the functional sites.

IV. Catalytically important groups and structure-forming amino acids in the active sites of enzymes

The answer to the question: which amino acids constitute active sites in enzymes the most often, was obtained by a comparison of amino acid sequences of proteins united in one large family.[11, 73, 74] The results of the alignment of sequences of large families of enzymes deposited with the database HSSP (http://www.sander.embl-heidelberg.de/hssp/) have been analysed.

The alignment is usually presented in the form of tables obtained by superposition of a protein sequence and a basic sequence. Visual detection of the conserved elements of a sequence is possible. Apparently, this method provides insufficient information where more than three to five proteins are to be compared. This can be automated to quantitate the conservation of amino acid positions in a sequence. The statistic criterion in the form of the Shannon entropy can serve as a quantitative measure of the conservation of each amino acid in the amino acid sequence of the protein.[75–77] It is noteworthy that the Shannon entropy is one of the most important functions in the theory of information. This was introduced as a measure of the degree of uncertainty which describes events characterised by a definite probability. In this case, the information can be defined as a measure of the level of uncertainty, which is corrected after the acquisition of information. Formally, the amount of information is presented as a difference between the information entropies before the experiment and upon its completion, *i.e.*, receipt of a message. The information entropy (the Shannon entropy) is the most convenient function for a comparison of related proteins differing in amino acid sequences. The alignment of sequences relative to a reference protein implies arrangement of their amino acid sequences one above the other with fixation of homologous sites, detection and elimination of the inserts. A comparison of a large panel of proteins allows calculation, with a sufficiently high degree of accuracy, of the probability of localisation of amino acids in each position of the protein sequence. This probability is defined as a relative frequency of the appearance of an amino acid of the type i in position j. The entropy function can be calculated for each position of all the 20 amino acids of the protein sequence.[75]

$$H_j = -\sum_i p_i^j \log_2 p_i^j.$$

An important feature of this function is that it is close to zero for both the events occurring with high ($p_i^j \to 1$) and low ($p_i^j \to 0$) probabilities. Calculation of the Shannon entropy for a protein sequence may result in the detection of common (*i.e.*, absolutely conserved) sites for the whole large family of proteins. This is a position where the probability of appearance of the given amino acid is close to unity, whereas that in the remaining positions is close to zero. High values of the Shannon entropy are characteristic of amino acids with variable positions in the amino acid sequences, being low for amino acids with conserved positions in the amino acid sequence. At the limit, for $p_i^j \to 1$ (absolute conservation), $H_j \to 0$ (Fig. 1).

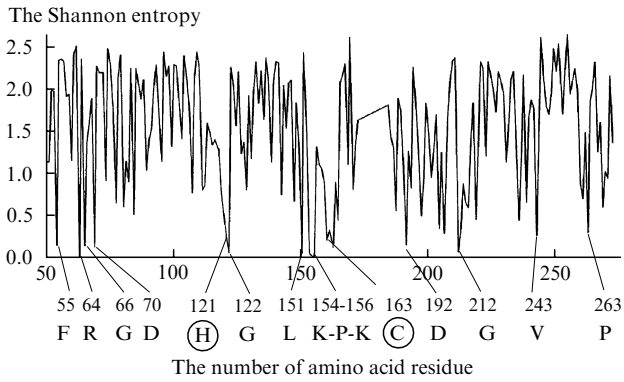


Figure 1. A presentation of alignment of the caspase family (PDB code 1cp3) in the form of the Shannon entropies. The amino acid residues of the active site are encircled.

The Shannon entropy criterion was used in the analysis of protein superfamilies with such basic proteins as trypsin/chymotrypsin, subtilisin, lipase, cytochrome C oxidase, pepsin/pepsinogen, enolase, lactate dehydrogenase, DNAse, aminoacyl-tRNA synthase and lysozyme.[73, 74] This sample of enzyme families comprises 2 281 proteins with the overall length of 249 252 amino acid residues.

This analysis revealed the following regularities.

1. Upon alignment of amino acid sequences in enzymes, the amino acids that form the catalytic site are always detected as conserved elements.

2. A comparison of amino acid sequences shows that glycine and aspartic acid are most commonly detected as absolutely conserved elements.

From our point of view, consideration of the results of the analysis of the most conserved amino acids with $H_j = 0$ (or close to zero) is of interest. A somewhat unexpected finding was that glycine appeared to be the most common conserved amino acid. The second position was occupied by aspartic acid; in total, glycine and aspartic acid constitute approximately one third of the conserved amino acids altogether.

The rating of amino acids according to their conservation in the protein families studied has been carried out. To this end, the frequency of occurrence of each amino acid as a conserved element ($H_j \approx 0$) was determined relative to the total number of conserved positions for all amino acids in the families. The conservation ratings of amino acids for two enzyme classes, *viz.*, hydrolases and oxidoreductases, are presented in Fig. 2. Despite the differences in the reactions catalysed by these enzymes, the mode of distribution of conserved amino acids virtually coincide.

It can be seen that glycine, aspartic acid, cysteine, proline, arginine and histidine in the amino acid sequences of enzymes usually manifest themselves as conserved amino acids and con-

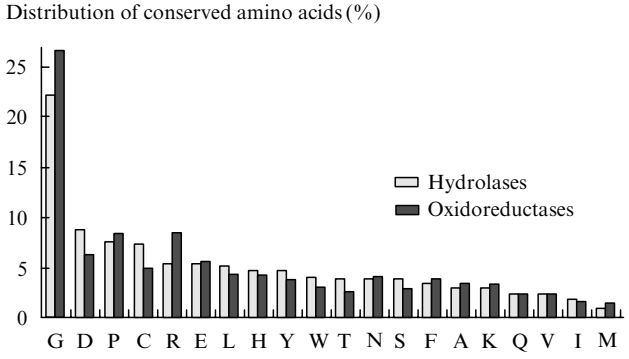


Figure 2. The rating of conservation of amino acids.

stitute ~ 60% of all conserved positions. Methionine and isoleucine are never conserved.

It seems logical to divide all conserved amino acids into two groups, different in principle, namely, amino acids involved in the elementary acts of activation of substrates as acids and bases (aspartic acid, histidine, arginine) and amino acids responsible for the architectonics of the active site (glycine, cysteine, proline).

A physicochemical interpretation of the role of acids, bases and structure-forming amino acids (glycine, proline, cysteine) in enzymatic catalysis will be given below in the example of hydrolases.

V. Classification of hydrolases according to the types of catalytic sites and catalytic mechanisms

In discussing the structures of active sites of enzymes, it is necessary to discriminate between two structural components of the active site, *viz*., a sorption subsite responsible for the binding, fixation and orientation of substrates (its properties determine the specificity of the enzyme),[1–3] and a catalytic subsite responsible for general acid-base catalysed chemical transformations of substrates.[4–6]

The structures and functions of hydrolases are extremely diverse. In addition to traditional involvement in the cleavage of biopolymers (proteins, nucleic acids, polysaccharides) into constituent monomers, hydrolases play a central role in some vital processes, such as regulation of fibrinolysis and thrombolysis,[78–80] apoptosis,[81–83] signal transmission in protein phosphorylation–dephosphorylation systems,[84–88] restriction and modification of DNA,[89,90] resistance against antibiotics,[91,92] regulation of the level of proteins,[93–95] nerve pulse transmission,[96] carcinogenesis,[97] *etc*. Within the framework of a traditional enzyme classification approved by the International Union of Biochemistry and Molecular Biology, hydrolases are classified according to the type of the reaction catalysed by them (hydrolysis) and enzyme specificity (the substrate structure). At present, the imperfection of this nomenclature is becoming more and more evident. On the one hand, the majority of enzymes do not possess absolute specificity and, which is even more important, manifest different types of activity. On the other hand, many enzymes with the same specificity and utilising the same substrate differ essentially in their primary structures, compositions of the active sites and fold of the polypeptide chain.

The analysis of the published data devoted to the mechanisms of action of different hydrolases suggests that despite the great diversity of enzymes, the number of catalytic sites that are different in principle is rather small.[73,74,98]

Hydrolases differ in the nature of the catalytic groups activating the water molecule in the course of the catalytic cycle. In principle, this information is imprinted in the spatial structure of a protein. The difficulty consists of an unambiguous identification of the catalytic group and the water molecule underwent activation. An active site may contain several water molecules among which only one is activated in the catalysis.

A comparative analysis of the compositions of active sites of hydrolases and mechanisms of activation of water revealed four main types of catalytic sites that are directly involved in the activation of water, *viz*.:

(1) hydrolases containing aspartic or glutamic acid in their active sites (pepsin and retroviral pepsins, glycosidases, dehalogenases and epoxide hydrolases);

(2) hydrolases containing histidine in their active sites (ribonucleases including pancreatic ribonuclease and barnase, acid phosphatases, hydrolases containing the Ser-His-Asp triad and its isomorphic substitutions, thiol hydrolases);

(3) zinc-dependent hydrolases containing one (carboxypeptidase A) or two (alkaline phosphatase, organophosphate hydrolase) zinc ion(s) used for the activation of water and the substrate;

(4) magnesium-dependent hydrolases (enzymes of the inorganic pyrophosphatase type, aryl sulfatase).

Taking into account the great abundance and significance of hydrolases containing serine, cysteine or threonine residues and a proton-transfer chain responsible for the activation of a hydroxy or a thiol group, it seemed expedient to group them in an individual subtype [the nucleophile–imidazole–aspartic (glutamic) acid subtype]. These enzymes form stable intermediates identified as acyl-enzymes in the course of catalysis.

The analysis of mechanisms of catalytic reactions shows that aspartic acid and histidine, in principle, play a crucial role in the majority of catalytic sites. The mechanisms of hydrolase catalysis can be illustrated by schemes that, though being rather approximate and not reflecting all the details of the reaction mechanisms, include their main stages, *viz*., activation of water and substrate hydrolysis.

1. Hydrolases containing aspartic or glutamic acid in their active sites

Hydrolases containing aspartic or glutamic acid in their active sites cleave ester, amide and glycosidic bonds. The catalytic sites of these enzymes contain two or more carboxy groups.[4,6] The non-dissociated carboxy group acts as an electrophile and takes part in the activation of the substrate as a generalised acid, while the deprotonated group acts as a nucleophile in the activation of the water molecule.[4,5] Enzymes of this group usually manifest maximum activity at relatively low values of pH. Pepsin (protease) and lysozyme (glycosidase) are typical representatives of this group.[99,100] The essential role of the carboxy groups of Asp-32 and Asp-215 in pepsin consists of formation of a hydrogen bond with each other and with the water molecule attacking the substrate (the spatial arrangement of the Asp-32 and Asp-215 residues is shown in Fig. 3). At optimum pH values, the carboxy group of Asp-32 becomes ionised, whereas the carboxy group of Asp-215 exists in the non-dissociated form. The processes occurring in the active site are shown in Scheme 1.[99–103]

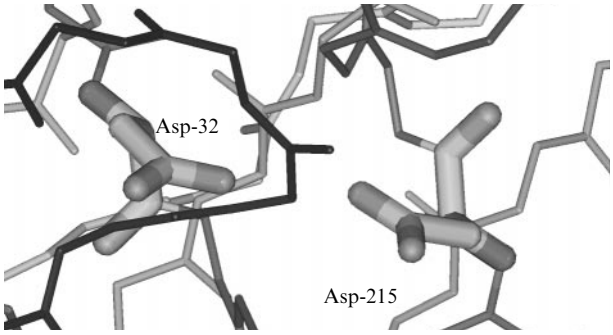
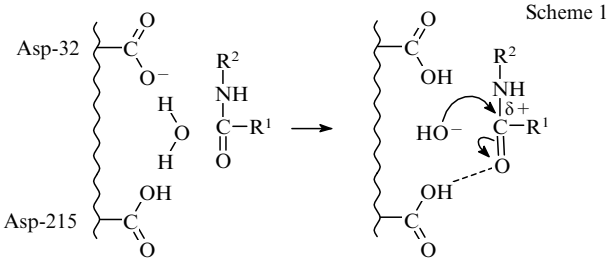


Figure 3. The active site of pepsin (PDB code 1flh). The residues of the active site are highlighted.

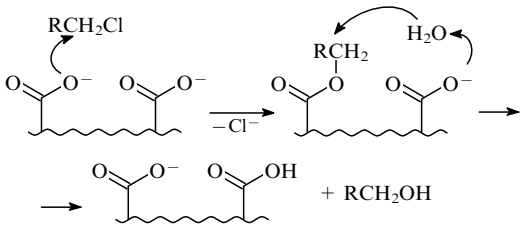
Scheme 1



Human immunodeficiency virus protease (retroviral pepsin) has the same mechanism of action, but its two aspartic acid residues (Asp-25) are localised in two identical subunits. It is assumed that animal pepsin was formed from a protein precursor similar to retroviral pepsin by duplication and concatenation of the corresponding genes. In the course of evolution, the local environments of two aspartic acid residues of pepsin changed, which resulted in the enhancement of its catalytic properties.[104]

The active sites of glycosidases also contain two aspartic (or glutamic) acid residues localised on opposite sides of the bond to be hydrolysed. One of these carboxy groups, *viz*., the non-ionised group, activates the substrate by a proton-transfer mechanism thus decreasing the electron density on the carbon atom of the bond to be hydrolysed. The second carboxy group activates the water molecule (the mechanism with inversion of configuration of the carbon atom) or attacks the substrate directly resulting in the formation of an intermediate which is further decomposed by the water molecule (the mechanism with retention of configuration).[105, 106] This mechanism operates in the case of lysozyme [107–109] and chitinases.[110, 111]

The three-dimensional structures of dehalogenases were established relatively recently.[112, 113] The interest in this group of enzymes is explained by the possibility of using them for the decomposition of halogen-containing organic environmental pollutants. The active sites of dehalogenases also contain two carboxy groups. One of them (a nucleophile) attacks the substrate with the elimination of the halide ion to give a covalent intermediate. The second carboxy group activates the water molecule involved in the dealkylation of the enzyme.



2. Hydrolases containing histidine residues in their active sites

Histidine residues of hydrolases are either involved in the activation of water or react with substrates directly.

According to classical concepts, the mechanism of catalytic cleavage of phosphodiester bonds in ribonucleic acids by pancreatic ribonuclease A (see Ref. 114) is provided by a concerted action of the imidazole groups of two histidine residues (Fig. 4). The imidazole group of His-12 activates the 2-hydroxy group of ribose by a proton transfer in the step involving the formation of a 2,3-cyclophosphate intermediate, whereas the imidazole group of His-119 activates the water molecule that hydrolyses the cyclophosphate.

The active site of barnase contains a glutamic acid residue (Glu-73) instead of His-12. A detailed mechanism of catalysis of barnase has been described.[115, 116]

In acid phosphatases,[117] histidine plays the role of a nucleophile that attacks the phosphorylated substrate. This results in the formation of phosphohistidine and the liberation of the non-substituted substrate (in the next step, the phosphoenzyme is hydrolysed with water).

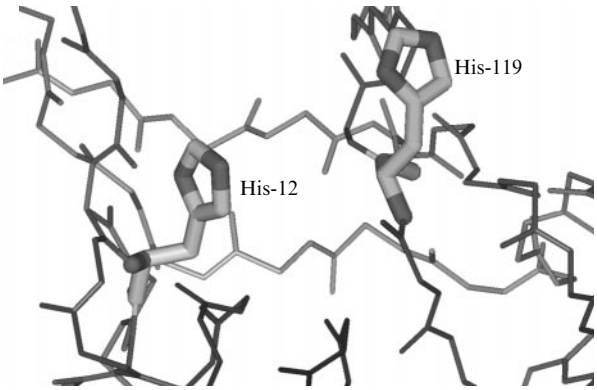


Figure 4. The active site of pancreatic ribonuclease (PDB code 1a5q).

It is of note that there is no hydrolase known at present for which the catalytic mechanism includes the formation of substituted histidine with the nitrogen of histidine and the carbon atom of the substrate. Apparently, this can be attributed to the fact that such a bond is relatively stable, so its cleavage would be a limiting step in the catalytic cycle.

3. Hydrolases containing a nucleophilic serine, cysteine or threonine residue

Serine proteases (chymotrypsin and trypsin) and cysteine protease (papaine) are typical representatives of this group of enzymes. Their catalytic mechanisms include acylation of the hydroxy or thiol group of the enzyme with the formation of an intermediate (acylenzyme) and its subsequent hydrolysis.[4, 118, 119] As the hydroxy group is a weak nucleophile, all the enzymes of this type contain amino acid residues essential for its activation and deprotonation (the proton-transfer chain). The same residues activate the water molecule in the enzyme deacylation step.

Hydrolases containing the catalytic triad Ser-His-Asp seem to be especially well studied. They include proteases (trypsin, chymotrypsin, elastase, subtilisin, carboxypeptidase Y and many others),[4, 120–123] lipase,[124] and acetylcholinesterase [125] (the latter contains a glutamic acid residue instead of aspartic acid). The active site of chymotrypsin contains Ser-195, His-57 and Asp-102 residues (Fig. 5). The catalytic mechanism of chymotrypsin is presented in Scheme 2.

Numerous isomorphic substitutions of the amino acids of the triad are known; this problem has been discussed in detail.[126] Thus thiol hydrolases contain a cysteine residue instead of serine. Since the thiol group is a stronger nucleophile than the hydroxy group, a shorter (in comparison with serine hydrolases) proton-transfer chain can be used. For example, many thiol proteases homologous

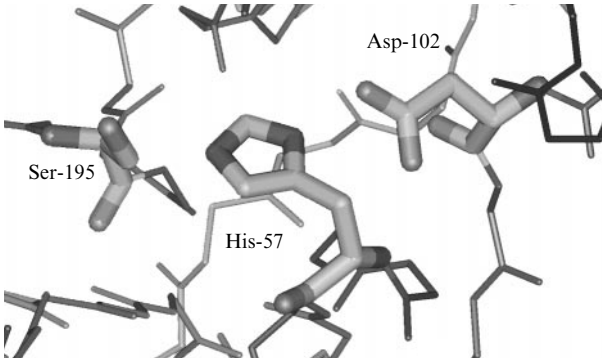
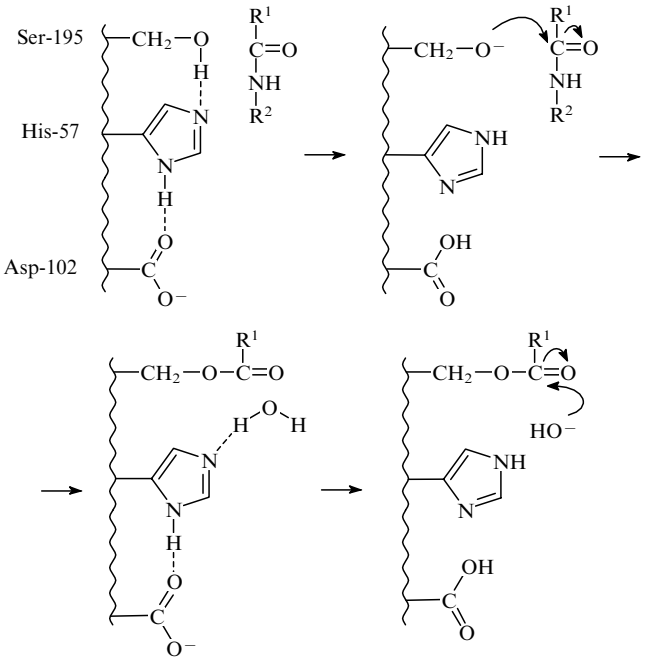


Figure 5. The active site of trypsin (PDB code 1fmg).

Scheme 2



to trypsin proteases contain asparagine able to form a hydrogen bond with histidine instead of aspartic acid, which polarises the imidazole ring of histidine. Some enzymes which do not involve histidine for the activation of the thiol group are known.[127]

Enzymes in which the amino group acts as a proton acceptor[128] in the activation of the hydroxy group constitute a separate group of enzymes. An example is a non-charged amino group in the side chain of lysine (β-lactamase[129] and asparaginase[130]). Serine, threonine or cysteine are N-terminal amino acids in N-terminal hydrolases, which include proteosome[131] and penicillinase components;[132] the free amino group in the main protein chain activates the hydroxy or the thiol group of the same amino acid.

4. Zinc(nickel, cobalt)-dependent hydrolases

This type of hydrolase contains one or several zinc ions (which can be replaced by cobalt or nickel without any loss of activity). Metal ions localised in the active sites of enzymes fulfil two main functions,[4,133,134] *viz*., that of an electrophilic agent activating the reaction centre of the substrate that is being attacked, and that of an electrophilic activator of water as a source of hydroxyl ions.

Enzymes of the carboxypeptidase A type, thermolysin or botulinic toxin,[135] each contains one zinc ion in their active sites; the role of ligands is played by aspartic or glutamic acid, histidine or cysteine residues. The metal ion complexes in the active sites of these enzymes act as electrophilic reagents which activate the reaction centres of the substrates. The activation of the attacking water molecule is induced by the deprotonated carboxy group of glutamic acid that is not involved in the coordination sphere of the zinc ion.

Enzymes of the type of alkaline phosphatase,[136–138] urease,[139] metal-dependent β-lactamase[92] or organophosphate hydrolase[140,141] contain two zinc ions in their active sites; these ions form structurally similar complexes with the protein (Fig. 6). Aspartic and glutamic acid residues as well as histidine (or cysteine) and N-carboxylysine residues act as metal-ion ligands. As a rule, one of the zinc ions activates the substrate by reducing electron density on the atoms being attacked, whereas the second zinc ion activates the water molecule or, in the case of alkaline phosphatase, the hydroxy group of serine. For organophosphate hydrolase,[140–144] the enzyme catalysing hydrolysis of organo-

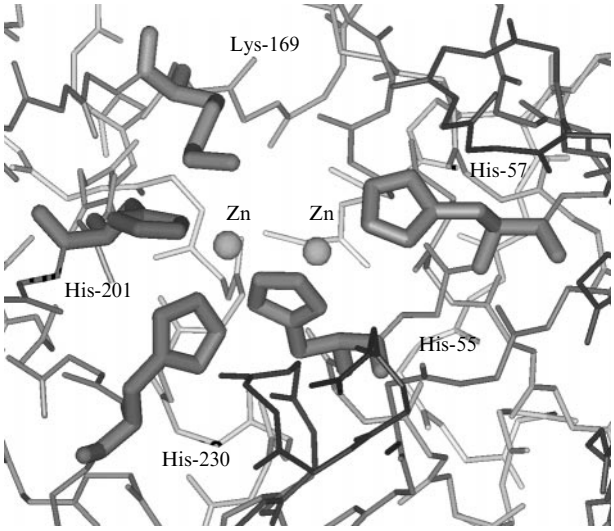
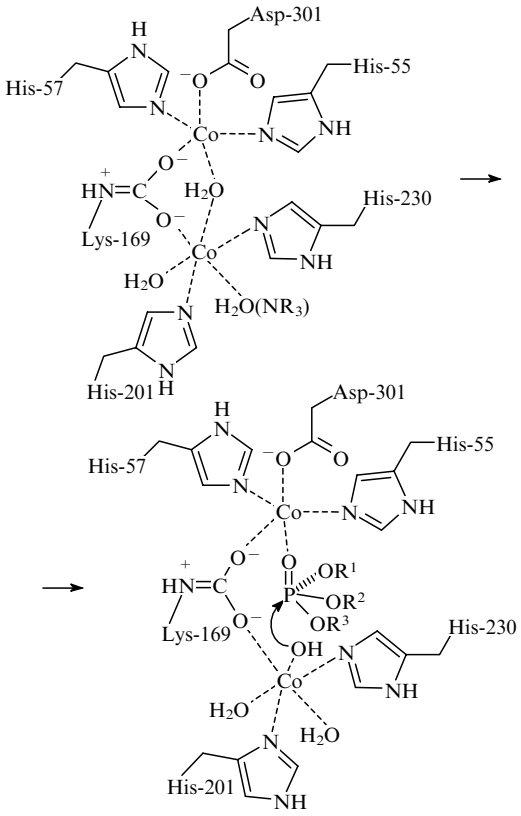


Figure 6. The active site of organophosphate hydrolase (PDB code 1dpm).

phosphorous pesticides and war gases such as Sarin, Soman or Vx, the catalytic mechanism[145] can be described by Scheme 3.

Scheme 3



5. Magnesium(manganese)-dependent hydrolases

This type of hydrolase contains magnesium or doubly charged manganese ions in the catalytic site.[146,147] As in the case of zinc-dependent enzymes, the metal complexes in the active sites of hydrolases behave as strong electrophilic reagents which activate the substrates (or the water molecule) by reducing electron density on the reaction centre. Such behaviour of metal ions became especially apparent in the analysis of the structure and catalytic mechanism of inorganic pyrophosphatase.[148–153] The substrate is fixed in the active site of the enzyme due to ionic interactions with

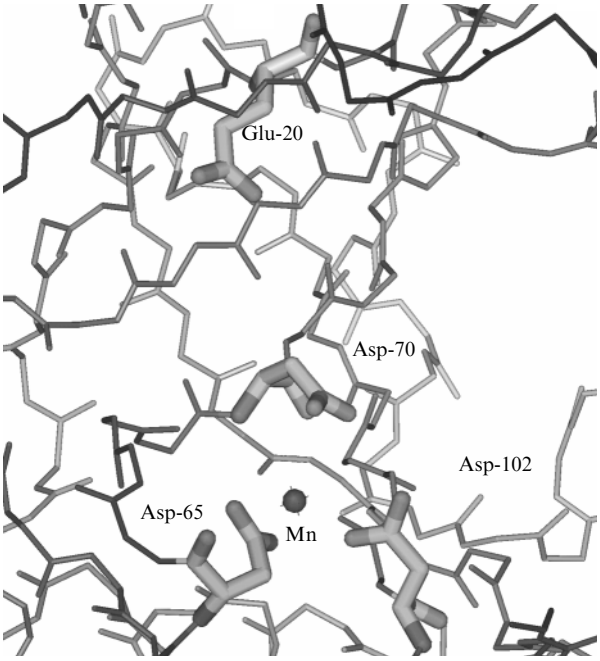
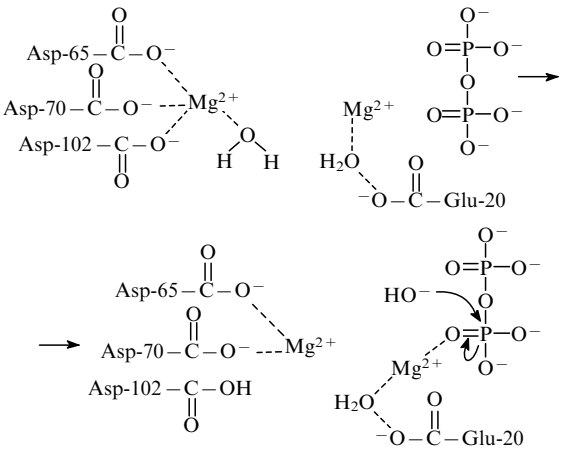


Figure 7. The active site of inorganic pyrophosphatase (PDB code 1ipw).

positively charged groups of the protein, Lys-103, Arg-78 or Lys-56. The main catalytic events involve two metal ions (Mg^{2+} or Mn^{2+}).[154] The carboxy groups of the residues Asp-65, Asp-70 and Asp-102 and three water molecules play the role of ligands for the ions in the high-affinity site where the formation of the Mg^{2+} (or Mn^{2+}) complex takes place. The low-affinity site is formed by the metal ion, water molecules and the carboxy group of the Glu-20 residue (Fig. 7). It can be assumed that the high-affinity site functions as the electrophilic activator of water (a generator of hydroxyl ions), while the low-affinity site plays the role of an electrophilic activator of the substrate.



These examples illustrate the role of aspartic acid and histidine in the mechanism of action of enzymes. Obviously, concerted action of nucleophilic and electrophilic components of the active site is the major principle in the functioning of active sites of enzymes.

It seems interesting to follow the transition from enzymes that catalyse hydrolysis of particular bonds to enzymes that catalyse the formation of these bonds. It is known that the active sites of exonucleases, which catalyse hydrolysis of DNA,[155–159] represent complexes of magnesium ions with the carboxy groups of aspartic and glutamic acids and resemble the active site of pyrophosphatase. The active site of DNA polymerase, which catalyses synthesis of DNA, is largely homologous to the active site of this nuclease.[155–159] The same functional groups of DNA polymerase activate the reaction centre and the nucleophile; however, the hydroxy group of the sugar (but not water) becomes activated owing to the abstraction of a proton. The nucleophilic attack on the reaction centre of the phosphate group ensures the growth of the DNA chain.

It should be noted that the structural elements constituting the active sites of hydrolases were also detected in the active sites of other classes of enzymes. Thus the proton-transfer chain, which includes the pair 'imidazole–carboxy group of aspartic acid' is often identified in the active sites of oxidoreductases. In peroxidases and in the protein environment of heme, the pair 'imidazole–carboxy group' plays the role of a generalised nucleophilic reagent which activates the iron ion of hemin.[160–162] The same structural element of the proton-transfer system, *viz*., the imidazole complex with the carboxy group, is identified in the active sites of dehydrogenases.[163–165]

VI. Glycosyl hydrolases, cellulases

The diversity of structures and fold of the polypeptide chain with simultaneous preservation of the main mechanism of water activation and the structure of the catalytic site can be illustrated in the examples of 1,4-β-glycosyl hydrolases and cellulases. Two carboxy groups of these enzymes, which induce concerted cleavage of 1,4-β-glycosidic bonds, are involved in the catalytic process.[106, 166] Xylanases, chitinases and lysozyme act in a similar way.[107–111] Cellulases are multidomain proteins, which contain one catalytic and one cellulose-binding domain. These enzymes are represented by a vast array of structures.[167] According to the fold of their polypeptide chains, they are classified into 15 clans, each containing > 90 structural families.[167–173] It is interesting to note that in contrast to 1,4-β-glycosidases, α-glycosidases are represented by only three families.[167, 174] The immense diversity of structural elements of catalytic domains of glycosyl hydrolases, of cellulases, in particular, is determined by the peculiarities of their primary sequences and the folding of their polypeptide chains. The database CAZY offers the most complete structural classification of glycosyl hydrolases. However, the structures of the catalytic sites of the whole array of proteins differing in their primary sequences and secondary structures are rather monotonous. The activation of the water molecule and the hydrolytic catalytic cycle are supported by the concerted action of two carboxy groups.

Enzymes containing two carboxy groups in their active sites, which represent concertedly acting electrophilic and nucleophilic agents, are especially widely occurring in nature. The results of alignment and calculation of the Shannon entropy for pepsin are shown in Fig. 8. As can be seen, Asp-32 and Asp-215 represent

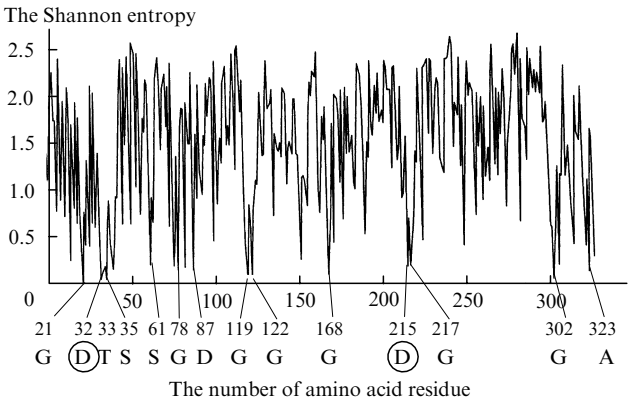


Figure 8. A presentation of alignment of the pepsin family (HSSP code 1pso) in the form of the Shannon entropy. The amino acid residues in the active site are encircled.

conserved amino acids. The activities of other proteases as well as peptidases, lysozymes, chitinases and many other hydrolases (the set of structures is virtually unlimited) are also mediated by the two carboxy groups. All these proteins are built in such a way as to favour the spatial approach and concerted action of two carboxy groups of the catalytic site. These data clearly demonstrate the structural paradox of enzymatic catalysis, namely: the immense diversity of protein structures provides the formation and assembly of a limited number of unique catalytic sites.[91, 175, 176]

VII. Serine hydrolases. Automatic search for a catalytic triad

We have carried out an automatic search for definite structures in the catalytic site of hydrolases based on an analysis of primary sequences within the framework of 'alignment' studies and calculations of the Shannon entropies using a template of three-dimensional structures of catalytically important residues. In the four classical types of active site structures of hydrolases, identification of catalytic groups of Zn(Co)- and Mg(Mn)-dependent enzymes is the least difficult. The localisation of the metal ion and its environment allows identification, with a high degree of reliability, of the components of the catalytic process. The identification of catalytic sites of hydrolases containing carboxy groups in the active sites was especially difficult because of the limited number of enzymes containing such catalytic sites and the lack of additional criteria characteristic of active sites and differentiating it from other potential structures, the use of an automatic procedure has led to the detection of a large number of false examples. Probably, this task would be solved more successfully in future upon introduction of additional structural limitations. The most adequate results were obtained in the analysis of protein structures containing the catalytic triad serine – histidine – carboxy group.

Serine hydrolases containing the catalytic triad Ser-His-Asp constitute the most well-studied representative group of hydrolases. This type of a catalytic site was first discovered in bovine pancreatic α-chymotrypsin (for a review see Ref. 177). Later, an identical catalytic site was found in bacterial subtilisin which is not homologous to chymotrypsin. The same triad was also found in lipases, while a similar triad (Ser-His-Glu) was detected in acetylcholinesterases. In more recent studies, new families of hydrolases (brain cortex esterases, virus envelope proteins, *etc.*) non-homologous to the known enzymes and containing the same residues in the catalytic site with the same spatial arrangement were established.[126] Various physiologically active substances, *e.g.*, cocaine [178] or the platelet activation factor,[179] can serve as the substrates.

Serine hydrolases form covalent intermediates with these substrates; in their active sites, the hydroxy group of the serine residue is acylated by the acidic fragment of the substrate (the acyl-enzyme-type intermediate). In the active sites of all enzymes of this type, the spatial localisation of the triad Ser-His-Asp is similar despite the difference in the fold of the polypeptide chain.

The catalytically important groups were identified using the method previously developed for estimating the conservation of residues deposited with the database HSSP, which is based on pairwise alignment and the Shannon entropies and employs the data on the coordinates of individual atoms used for data retrieval in the structural bases of these proteins.

The data on the arrangement of residues in chymotrypsin and subtilisin (for typical arrangement of the residues see Fig. 9) were used for designing a template for the catalytic triad (the constraints on type of the appropriate amino acid residues are listed in Table 1); this was followed by a search on the set of non-homologous hydrolases from the data bank PDB. Over 100 structures with a required arrangement of catalytically important functional groups containing unique primary sequences were found.

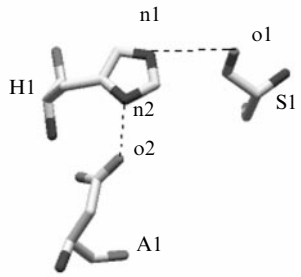


Figure 9. A typical distribution of catalytic triad residues. The names of amino acid residues are capitalised; the names of individual atoms are given in lower case letters. The dotted lines indicate the distances between the atoms on which certain restrictions were imposed in the search (for explanations see text).

Table 1. The criteria of a search for the triads Ser-His-Asp through the protein structures.

Type of limitation	Parameter [a]	Allowable values, amino acids, atoms
Conservation of the amino acid residue	value of the Shannon entropy	< 0.9
Limitation for the type of the amino acid residue	residue S1	Ser
	residue H1	His
	residue A1	Asp or Glu
Limitation for the type of the atom	atom o1	atom OG [b] of residue S1
	atom n1	atom NE2 [b] of residue H1
	atom n2	atom ND1 [b] of residue H1
	atom o2	atom OD1, OD2, OE1 or OE2 [b] of residue A1
Limitation for the distance between atoms	distance between atoms o1 and n1	⩽4 Å
	distance between atoms n2 and o2	⩽3.4 Å
Limitation for allowable angles between three atoms	angle between atoms o1, n1, n2	100° – 180°
	angle between atoms n1, n2, o2	125° – 180°

[a] The designations of the atoms and amino acid residues coincide with those used in Fig. 9. [b] The names of the atoms correspond to the nomenclature of the Protein Data Bank.

The next step included the use of the database SCOP, where the proteins are systematised according to the fold of the polypeptide chain (see above). The data on the folding of the catalytic domain for each protein were retrieved from the database SCOP. The two upper levels of the hierarchy obtained are shown in Fig. 10. The most characteristic enzymes of each group are listed in Table 2. It was found that all serine hydrolases containing the triad Ser-His-Asp(Glu) are made up exclusively of β-sheets (*e.g.*, chymotrypsin), or irregularly alternating α-helices and β-sheets (*e.g.*, subtilisin). None of the hydrolases of this type, known so far, are built up exclusively of α-helices. Considering that primary sequences differ in the types of spatial folding of the peptide chain, all hydrolases can be divided into families according to this particular criterion. By choosing different levels of folding, hydrolases can be divided into families with different levels of similarity. From Fig. 10 it follows that all Ser-His-Asp-containing hydrolases form two clans; it is noteworthy that the clan of proteins built up of β-sheets comprises only one family, while the clan of α,β-proteins includes five different families. These families differ not only in the type of chain folding, but also in the order of

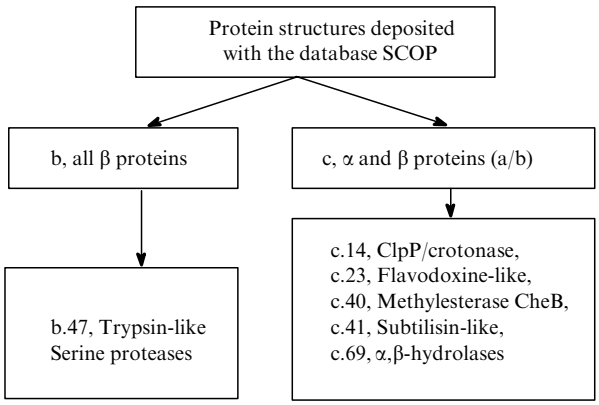


Figure 10. The classification of protein structures deposited with the database SCOP.

Table 2. The order of arrangement of amino acid residues in a catalytic centre and typical representatives of the serine hydrolase family containing the catalytic triad Ser-His-Asp, and classified according to the fold of the peptide chain.

Type of folding of the polypeptyide chain	Triads [a]	Typical representatives of the given class
b.47, Trypsin-like	His-Asp-Ser	α-tryptase, viral proteases, collagenase, a complement component plasminogen, mast cell protease, prothrombins, strepto-grisin, trypsin, blood coagulation factors, chymotrypsin, elastase
c.14, ClpP/crotonase	Ser-His-Asp	ATP-dependent Clp-protease
c.23, Flavodoxin-like	Ser-Asp-His	acyl-CoA-thioesterase, platelet activation factor
c.40 Methylesterase CheB	Ser-His-Asp	methyl glutamate hydrolase
c.41, Subtilisin-like	Asp-His-Ser	kexin, subtilisin, furin
c.69, α,β-Hydrolases	Ser-Asp(Glu)-His	acetylcholinesterase, lipase, cocaine esterase

[a] The order of the distribution of the catalytic site residues in the primary protein sequence.

distribution of active site residues in the primary sequence (see Table 2).

It is of note that these results were obtained as a result of an automatic search for proteins in the structural database PDB and were confirmed by independent literature data. Thus, the approach based on the use of automatic bioinformatics methods and databases in the information selection step permits the processing of the tremendous bulk of accessible information.

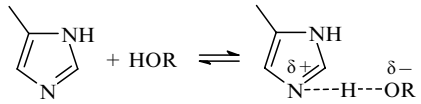
VIII. Mechanism of activation of substrates by active sites of hydrolases

Acids and bases of the active sites play a crucial role in enzymatic catalysis by activating the substrates. Aspartic acid and histidine are most often involved in the activation of molecules in the enzymatic catalytic cycles in comparison with other amino acids. Presumably, the catalytic function of aspartic acid and histidine in the active sites of enzymes is determined by their properties as an acid and a base.[4, 5, 11]

In the non-dissociated form, the carboxy group of aspartic acid represents a proton donor, which manifests pronounced electrophilic properties, whereas in the deprotonated form it plays the role of a proton acceptor (a nucleophilic agent), *viz*., an electron density donor for the corresponding reaction centre. The imidazole group of histidine is a pronounced nucleophilic reagent able to donate unshared electron pairs of nitrogen atoms.

Several schemes of substrate activation by histidine and aspartic acid residues represent the main reactions occurring in the catalytic site.

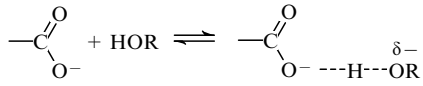
1. Nucleophilic activation of the substrate by the imidazole group.



(1)

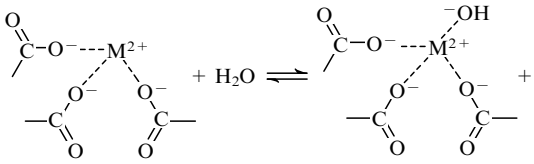
In this case, a negative charge is generated in the reaction centre.

2. Nucleophilic activation of the substrate by the carboxy group of aspartic acid.



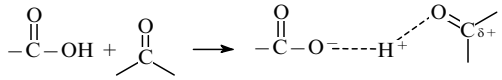
(2)

3. Electrophilic activation of water. Metal ions (Zn^{2+}, Co^{2+}, Ni^{2+}, Mg^{2+}, Mn^{2+}) present in the active sites of enzymes as complexes with the carboxy groups of aspartic or glutamic acid can take part in the catalytic cycle as both acids and bases. Metal ions can replace the proton in the molecule of water thus inducing a highly reactive and mobile hydroxyl ion.

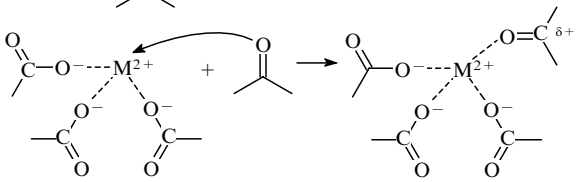


(3)

4. Electrophilic activation of the reaction centre. The electrophilic activation of the reaction centre of the substrate (*e.g*., the carbonyl group) can be induced by both the carboxy group and the metal ion.



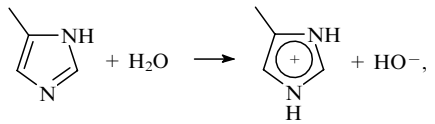
(4)



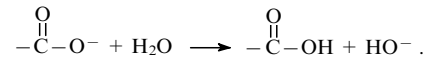
(5)

These reactions reduce the electron density on the reaction centre.

For hydrolytic reactions, the water activation steps in the nucleophilic catalysis can be presented as follows:



(6)



The reactivity of the water molecule increases 10^7-fold upon its conversion into hydroxyl ions.[5] Assuming that the electrophilic activation of the reaction centre, which proceeds according to the mechanisms (4) and (5), increases its reactivity 10^6-10^7-fold, the high efficiency of the enzymatic catalysis and its ability to increase the reaction rate $10^{13}-10^{14}$-fold can be explained in the first approximation. The hypothesis that acids and bases act concertedly in the catalysis was put forward on the basis of an analysis of some independent studies.[5, 180, 181]

The degree of charge transfer in various activation processes occurring in accordance with Schemes (1)–(6) in the transition states of enzymatic reactions is still questionable.[182] It is obvious, however, that the activation of well-oriented fixed (relative to one another) molecules takes place in the reaction centres. Of crucial importance is the fact that acids and bases within the catalytic site form a system of hydrogen bonds with one another, which opens up unique opportunities for the proton transfer. In systems containing hydrogen bonds, acids and bases react much faster than in solution. The interactions of hydroxyl ions with protons in a system of water molecules structured by hydrogen bonds in ice crystals (studied by Eigen [183]) provides an illustrative example. It was found that the reaction rates of the hydroxyl ion with hydrogen ion exceed those of diffusion-controlled processes by 3–4 orders of magnitude. As a matter of fact, the proton transfer and charge neutralisation are complete within a single vibration with a frequency of 10^{14} s^{-1}. The precise orientation of reactants and catalytic site components is a salient feature of enzymatic catalysis.[184]

Identification of acids and bases (electrophilic and nucleophilic reagents) within the catalytic site is the primary goal in the elucidation the mechanisms involved in the functioning of catalytic sites of enzymes. The activation mechanisms presented in Schemes (1)–(6) provide an explanation for the high efficiencies of the majority of presently known enzymatic reactions. We described an algorithm for identification of the active groups of the catalytic site. The knowledge of amino acid sequences of the protein allows determination of a group of proteins (enzymes) homologous to the given protein using bioinformatics methods. The alignment procedure and calculation of the Shannon entropies permits one to determine the positions of conserved amino acids. The latter represent, with a high degree of probability, catalytically important residues, which ensure substrate activation and, correspondingly, determine the mechanism of catalytic action.

IX. Glycine, cysteine and proline as structure-forming amino acids

A comparison of amino acid sequences in homologous families showed that conserved glycine, cysteine and proline residues are essential elements in the active sites of enzymes. The role of glycine in the formation and functioning of active sites is not so apparent as the role of aspartic acid. Evidently, conserved glycine residues do not play any pronounced role in the activation reactions in the catalytic cycle. Nevertheless, this amino acid is a very important structural element of the protein [73, 74, 185–190] as can be evidenced from the results of its site-specific substitution for any other amino acid. As a rule, such a substitution results in a complete loss or a significant decrease of the enzyme activity.[188]

Studies of the substrate specificity of the functional domain of halogenocarboxylic acid dehydrogenases [191] revealed that mutations in the conserved Gly residue localised between the amino acid residues of the catalytic site to Pro, Val or Ala reduce dramatically the catalytic activity. Structural studies of mutants have demonstrated substantial conformational changes in the side chains of the residues forming the catalytic site. It was concluded that the presence of Gly is critical for conferring flexibility on the polypeptide chain in its folding in the vicinity of the active site. The role of conserved Gly residues in the functioning of gramicidin S synthase has been studied.[192] It was shown that single mutations in three conserved Gly residues to Glu or Asp result in the loss of the valine-adenylate activity of this enzyme.

Apparently, conserved glycine residues are important in principle for the following reasons. First, glycine, being a unique amino acid with the smallest energy of rotation around the C—N and C—C bonds of the polypeptide chain plays the role of a node which determines the direction of the folding of the polypeptide chain in the assembly of the protein globule. Thus, the presence of conserved glycine residues accounts for the structural paradox of enzymatic catalysis, *viz*., the 'assembly' of identical active sites from absolutely different polypeptide chains. The presence of conserved glycine residues and stabilisation factors in the 'assembled' structures, *e.g*., disulfide bonds, are a common feature of these chains.

The substantial effect of conserved glycine residues on structure formation of the active sites of enzymes was established in computer studies.[74] Computer simulation was used to study the changes in the conformations of amino acid residues in the active sites of different types of hydrolases. Two series of calculations were carried out for each representative of the hydrolase family. In the first series, all conserved Gly were replaced by Ala and the corresponding conformational changes in the polypeptide chain in the vicinity of the active site were determined. Similar calculations were performed in the second series except that the non-conserved Gly residues were replaced by Ala. The root-mean-square deviation of the coordinates of the atoms in active site residues (calculated per one substituted amino acid residue) in the structure calculated after the substitutions from the coordinates in the active site of the native enzyme following superposition of these structures was used as a measure of distortion of the geometry of the active sites of these enzymes. It was found that the substitution of Ala for conserved Gly for all types of hydrolase under study caused more significant distortions in the geometry of the active sites of enzymes than the substitution for a greater number of non-conserved Gly. This suggests enhanced 'sensitivity' of the geometry of the active sites of enzymes to the presence of Gly residues in conserved positions.

Second, conserved glycine residues can play the role of conformational 'hinges' which impart conformational mobility to the catalytic groups. Supporting evidence can be derived from the fact that very often the conserved glycine residues are found in the vicinity of catalytically active groups. For example, the conserved motifs for hydrolases belonging to different families are as follows: Asp-215-X-Gly-217 (pepsin); Asp-170-X-X-Gly-173 (thermolysin); Asp-32-X-Gly-34, His-63-Gly-64, Gly-219-X-Ser-221 (subtilisin); Gly-173-X-Ser-177 (trypsin); His-76-Gly-77, Ser-153-X-Gly-155, Gly-175-X-Asp-177, Gly-X-Ser-X-Gly (lipases).[193] The residues Asp, Ser and His of these enzymes constitute the active sites.

Enzyme-catalysed transformation of the starting substrate into final products is accompanied by the formation of many intermediates with structures different from that of the starting substrate. Glycine residues of the active site can play the role of 'relaxing' elements by tuning the conformation of the active site for the subsequent elementary act.

A prominent role in the formation of the active site architecture is played by cysteine and proline residues (the fourth and third positions in the rating list of conserved amino acids). Proline is known to be a unique amino acid responsible for the unfolding of the polypeptide chain.

In all probability, the role of cysteine residues consists in the 'fixing' of the required configuration of the active site built up of different, often remote fragments of the polypeptide chain, by forming a chemical bond in the form of a disulfide bridge. For many enzymes, this process completes the formation of the active site architecture.

X. The rates of conformational changes and proton transfer as the limiting steps

The numeric values of the rate constants (k_{cat}) of the limiting and known elementary steps of catalytic processes involving enzymes are worth analysing and comparing. To date, much is known on V_{max}, K_m (the Michaelis constant) and k_{cat} values.

Figure 11 shows the distribution density curves of different enzymes according to their k_{cat}, K_m and k_{cat}/K_m values and the rate constants for the elementary steps (k_1, k_{-1} and k_{conf}) (see Ref. 11). The shapes of these curves are characteristic of normal logarithmic distribution.

The k_{cat} values lie in the range of $10^1 - 10^4$ s^{-1}. Enzymatic reactions in which the limiting step is of the order of 10^2 s^{-1} are especially abundant in nature. The distribution patterns of enzymes according to their K_m values are similar. This parameter changes in the range from 10^{-2} to 10^{-7} mol litre^{-1}. The distribution density has a maximum at $K_m \approx 10^{-4}$ mol litre^{-1}. The k_{cat}/K_m values for different enzymes lie in the range of $10^1 - 10^{10}$ litre mol^{-1} s^{-1}; enzymes with $k_{cat}/K_m = 10^6$ litre mol^{-1} s^{-1} are especially numerous.

These values define the limiting step of the substrate conversion in the active site of the enzyme. The energy barrier (10–17 kcal mol^{-1}) seems to be determined by physicochemical factors. Other reaction steps proceed at faster rates.

Structural isomerisations are substantial steps of the catalytic cycle. If conformational changes affect the active site of the enzyme, this is reflected in the kinetics of the enzymatic reaction. In many cases, binding of the enzyme to the substrate (or the inhibitor) occurs by a two-step mechanism:

$$E + S \rightleftharpoons X_1 \rightleftharpoons X_2,$$

where E is the enzyme, S is the substrate, X_1 is the primary complex and X_2 is the enzyme–substrate (or enzyme–inhibitor) complex.

As the formation of enzyme–substrate (or enzyme–inhibitor) complexes is not accompanied by the formation of covalent bonds, the second step of the reaction represents a conformational transformation of the primary complex X_1. Presumably, this fact provides the strongest evidence in favour of conformational changes in the active site of the enzyme during enzymatic catalysis.[11]

Very often, the conformational changes $X_1 \rightleftharpoons X_2$ occur at faster rates ($10^3 - 10^4$ s^{-1}) than the limiting step of the catalytic process.

The rate constants of many conformational changes in proteins involved in catalytic transformations have now been established. These data can be used for conducting statistical analyses and revealing 'typical' characteristics of the catalytic reaction. Figure 11 *a* shows the curve of the distribution density of rate constants of conformational changes (k_{conf}). The maximum of the distribution density lies around 10^3 s^{-1}. It can be seen that the distribution density curve k_{conf} is very similar to that for the limiting steps. It may thus be assumed that sometimes the rates of conformational changes determine the rates of enzymatic reactions.

However, in the majority of cases conformational changes occur faster and precede the rate-limiting step. The role of conformational changes in enzymatic catalysis is evident. The mechanisms of enzymatic reactions are rather complex involving rearrangement of numerous bonds. Such processes cannot occur in one elementary step. The mechanisms of action of enzymes include the formation and conversion of a great number of intermediates. The enzyme ensures the reaction pathway with the least energy and entropy expenditures. The mechanism of action of an enzyme can be compared to that of a programmed control machine which operates at the molecular level and its functioning is provided by conformational changes.

In terms of chemical kinetics, this means that the structure similar to the transition state is formed as a result of conformational changes in the ground state. Conformational changes 'push' the reactants along the reaction pathway. The formation of a conformer structurally similar to the transition state or the reaction product significantly lowers the potential reaction barrier. The lowering of the reaction barrier the greater the more similar the structure of the conformer to that of the transition state or the reaction product.

Proton transfer is yet another process which limits the rate of an enzymatic reaction. It follows from the above-said that the acid-base nature of enzymatic reactions is their main characteristic feature.

Owing to the classical works by Eigen,[194, 195] the mechanisms of reactions of the proton with bases and of the hydroxyl ion with acids are now well understood. The main conclusion is that these reactions do not depend on the basicity or acidity of the reactants, but represent diffusion-controlled processes ($k \approx 10^{10}$ litre mol^{-1} s^{-1}). Correspondingly, the rate constants of reverse reactions are determined by the values of equilibrium constants. In some cases, the proton transfer steps can be limiting for the enzymatic process.

Since in the course of an enzymatic reaction the catalytically active group undergoes both protonation and deprotonation, the proton transfer poses certain restrictions on the efficiency of acid-base catalysis. The reactions occurring in this process can be described as:

$$BH^+ \rightleftharpoons B + H^+, \quad (7a)$$

$$B + H_2O \rightleftharpoons BH^+ + OH^-. \quad (7b)$$

Since the reverse steps of these reactions are controlled by the diffusion of the reactants, the rate constants of the direct reactions can be presented as $10^{10}K_a$ s^{-1} and $10^{10}(K_W/K_a)$ litre mol^{-1} s^{-1} (K_a is the acid dissociation constant of BH^+ and K_W is the ionisation constant of water). The maximum value of the rate constant of the overall reaction is approximately equal to 10^3 s^{-1} at $K_a \approx 10^{-1}$ mol litre^{-1}. Indeed, for $K_a < 10^{-7}$ mol litre^{-1}, the dissociation rate constant of the acid BH^+ [reaction (7a)] is

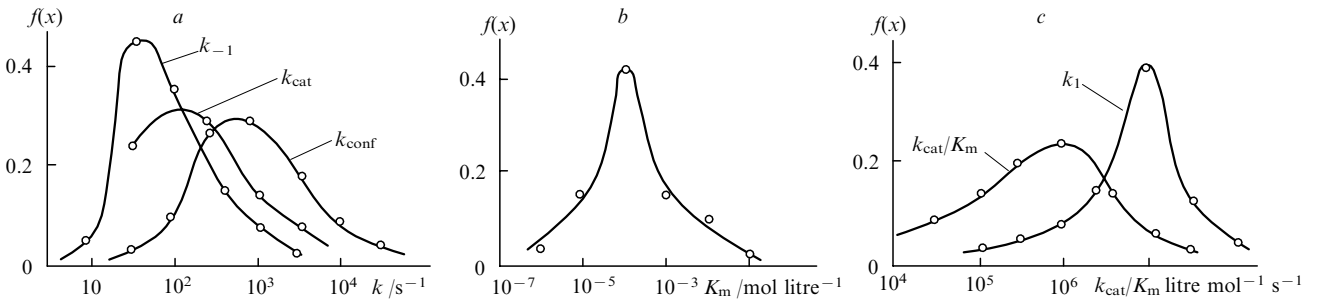


Figure 11. The distribution of enzymes according to their k_{cat} (*a*), K_m (*b*) and k_{cat}/K_m (*c*). values.
(*a*) First-order rate constant, (*b*) the Michaelis constant, (*c*) second-order rate constant.

smaller than 10^3 s^{-1}, while at $K_a > 10^{-7}$ mol litre^{-1}, the hydrolysis rate constant of the base [reaction (7b)] is smaller than 10^3 s^{-1}. Therefore, the maximum value of the rate constant for an acid-base catalysis involving water is equal to 10^3 s^{-1}. This conclusion can be used for interpreting the maximum rates of enzymatic catalysis. The value 10^3 s^{-1} is characteristic of the turnover number of many enzymes (see Fig. 11 *a*). Besides, the optimum value of $pK_a \approx 7$, which is necessary for effective catalysis, corresponds to pK_a of the imidazole group of the histidine residue.

Analysis of the structures of active sites of enzymes shows that in many cases the imidazole group of histidine participates in the catalytic act as a complex with a stronger nucleophile (proton conjugation), *viz*., a deprotonated carboxy group of aspartic or glutamic acid. The kinetic significance of such conjugation is obvious. The proton transfer from the protonated imidazole to the carboxy group in the last step of the catalytic cycle enables regeneration of properties of the catalytic site within microseconds. Indeed, the rate constant for the reaction

$$R-COOH \longrightarrow R-COO^- + H^+$$

is equal to 10^6 s^{-1}.

Thus, the enzymatic catalysis illustrates the maximum catalytic accelerations which can be realised for the reactions occurring in solutions.

In all probability, the evolution of biocatalysis has long been searching for effective structures by the trial-and-error method. Having fixed these structures at the genetic level, the biosystems have learnt to use them for the formation of various active sites, and not only of the hydrolase type.

References

1. I V Berezin, K Martinek *Osnovy Fizicheskoi Khimii Fermentativnogo Kataliza* (Foundation of The Physical Chemistry of Enzymatic Catalysis) (Moscow: Vysshaya Shkola, 1977)
2. S D Varfolomeev, S V Zaitsev *Kineticheskie Metody v Biokhimicheskikh Issledovaniyakh* (Kinetic Methods in Biochemical Investigations) (Moscow: Moscow State University, 1982)
3. A A Klesov *Uspekhi Biologicheskogo Kataliza* (Progress in Biological Catalysis) (Moscow: Moscow State University, 1979)
4. V K Antonov *Khimiya Proteoliza* (The Chemistry of Proteolysis) (Moscow: Nauka, 1991)
5. W P Jencks *Catalysis in Chemistry and Enzymology* (New York, London, Mexico: McGraw Hill, 1969)
6. M Dixon, E C Webb *Enzymes* (3rd Ed.) (London: Longman, 1979)
7. J R Wild, S D Varfolomeev, A Scozzafava (Eds) *Perspectives in Bioremediation Technologies for Environment Protection* (Dordrecht, Boston, London: Kluwer Academic, 1997)
8. B Zwanenburg, M Mikolajczyk, P Klielbasinski (Eds) *Enzymes in Action. Green Solutions for Chemical Problems* (New York: Kluwer Academic, 2001)
9. K Faber *Biotransformation in Organic Chemistry* (3rd Ed.) (Heidelberg: Springer, 1997)
10. U T Bounscheuer, R J Kazlauskas *Hydrolases in Organic Synthesis. Regio and Stereoselective Biotransformations* (Weinheim: Wiley, 1999)
11. S D Varfolomeev *Khimicheskaya Enzimologiya* (Chemical Enzymology) (Moscow: Akademiya, 2004)
12. D A Kraut, K S Carroll, D Herschlag *Annu. Rev. Biochem.* **72** 517 (2003)
13. J Jeffery *EXS* **73** 79 (1995)
14. B Boeckmann, A Bairoch, R Apweiler, M-C Blatter, A Estreicher, E Gasteiger, M J Martin, K Michoud, C O'Donovan, I Phan, S Pilbout, M Schneider *Nucleic Acids Res.* **31** 365 (2003)
15. R Apweiler, A Bairoch, C H Wu, W C Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, M J Martin, D A Natale, C O'Donovan, N Redaschi, L S Yeh *Nucleic Acids Res.* **32** 115 (2004)
16. T Kulikova, P Aldebert, N Althorpe, W Baker, K Bates, P Browne, A van den Broek, G Cochrane, K Duggan, R Eberhardt, N Faruque, M Garcia-Pastor, N Harte, C Kanz, R Leinonen, Q Lin, V Lombard, R Lopez, R Mancuso, M McHale, F Nardone, V Silventoinen, P Stoehr, G Stoesser, M A Tuli, K Tzouvara, R Vaughan, D Wu, W Zhu, R Apweiler *Nucleic Acids Res.* **32** 27 (2004)
17. A Bateman, L Coin, R Durbin, R D Finn, V Hollich, S Griffiths-Jones, A Khanna, M Marshall, S Moxon, E L Sonnhammer, D J Studholme, C Yeats, S R Eddy *Nucleic Acids Res.* **32** 138 (2004)
18. H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, P E Bourne *Nucleic Acids Res.* **28** 235 (2000)
19. D Baker, A Sali *Science* **294** 93 (2001)
20. M Garcia-Viloca, J Gao, M Karplus, D G Truhlar *Science* **303** 186 (2004)
21. R A Laskowski, E G Hutchinson, A D Michie, A C Wallace, M L Jones, J M Thornton *Trends Biochem. Sci.* **22** 488 (1997)
22. N M Luscombe, R A Laskowski, D R Westhead, D Milburn, S Jones, M Karmirantzou, J M Thornton *Acta Crystallogr., Sect. D* **54** 1132 (1998)
23. R A Laskowski *Nucleic Acids Res.* **29** 221 (2001)
24. C A Orengo, A D Michie, S Jones, D T Jones, M B Swindells, J M Thornton *Structure* **5** 1093 (1997)
25. F M G Pearl, D Lee, J E Bray, I Sillitoe, A E Todd, A P Harrison, J M Thornton, C A Orengo *Nucleic Acids Res.* **28** 277 (2000)
26. A G Murzin, S E Brenner, T Hubbard, C Chothia *J. Mol. Biol.* **247** 536 (1995)
27. L L Conte, S E Brenner, T J P Hubbard, C Chothia, A G Murzin *Nucleic Acids Res.* **30** 264 (2002)
28. A Andreeva, D Howorth, S E Brenner, T J P Hubbard, C Chothia, A G Murzin *Nucleic Acids Res.* **32** 226 (2004)
29. R Durbin, S Eddy, A Krogh, G Mitchison *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge: Cambridge University Press, 1998)
30. J Goug, K Karplus, R Hughey, C Chothia *J. Mol. Biol.* **313** 903 (2001)
31. J E Butler-Ransohoff, S E Rokita, D A Kendall, J A Banson, K S Carano, E T Kaiser, A R Matlin *J. Org. Chem.* **57** 142 (1992)
32. B Stec, M J Hehir, C Brennan, M Nolte, E R Kantrowitz *J. Mol. Biol.* **277** 647 (1998)
33. R Han, J E Coleman *Biochemistry* **34** 4238 (1995)
34. J E Butler-Ransohoff, D A Kendall, E T Kaiser *Proc. Natl. Acad. Sci. U.S.A.* **85** 4276 (1988)
35. A Chaidaroglou, D J Brezinski, S A Middleton, E R Kantrowitz *Biochemistry* **27** 8338 (1988)
36. A Matagne, J M Frere *Biochim. Biophys. Acta* **1246** 109 (1995)
37. J D Otvos, J R Alger, J E Coleman, I M Armitage *J. Biol. Chem.* **254** 1778 (1979)
38. P Gettins, J E Coleman *J. Biol. Chem.* **258** 408 (1983)
39. P Gettins, J E Coleman *J. Biol. Chem.* **259** 4991 (1984)
40. P Gettins, M Metzler, J E Coleman *J. Biol. Chem.* **260** 2875 (1985)
41. N D Rawlings, D P Tolle, A J Barrett *Nucleic Acids Res.* **32** 160 (2004)
42. N D Rawlings, A J Barrett *Biochem. J.* **290** 205 (1993)
43. N D Rawlings, A J Barrett *Methods Enzymol.* **244** 19 (1994)
44. N D Rawlings, A J Barrett *Methods Enzymol.* **244** 461 (1994)
45. N D Rawlings, A J Barrett *Methods Enzymol.* **248** 105 (1995)
46. N D Rawlings, A J Barrett *Arch. Biochem. Biophys.* **318** 247 (1995)
47. N D Rawlings, A J Barrett, in *Proteolysis in Cell Functions* (Eds V K Hopsu-Havu, M Jarvinen, H Kirschke) (Amsterdam: IOS Press, 1997) p. 13
48. A J Barrett, N D Rawlings, J F Woessner *Handbook of Proteolytic Enzymes* (London: Academic Press, 1998)
49. J-M Chen, N D Rawlings, A J Barrett *FEBS Lett.* **441** 361 (1998)
50. A J Barrett, N D Rawlings, E A O'Brien *J. Struct. Biol.* **134** 95 (2001)

51. P M Coutinho, B Henrissat, in *Recent Advances in Carbohydrate Bioengineering* (Eds H J Gilbert, G Davies, B Henrissat, B Svensson) (Cambridge: Royal Society of Chemistry, 1999) p. 3
52. P M Coutinho, B Henrissat, in *Genetics, Biochemistry and Ecology of Cellulose Degradation* (Eds K Ohmiya, K Hayashi, K Sakka, Y Kobayashi, S Karita, T Kimura) (Tokyo: Uni Publishers Co., 1999) p. 15
53. B Henrissat *Biochem. J.* **280** 309 (1991)
54. B Henrissat, A Bairoch *Biochem. J.* **293** 781 (1993)
55. B Henrissat, A Bairoch *Biochem. J.* **316** 695 (1996)
56. G Davies, B Henrissat *Structure* **3** 853 (1995)
57. M L Sinnott *Chem. Rev.* **90** 1171 (1990)
58. J A Campbell, G J Davies, V Bulone, B Henrissat *Biochem. J.* **326** 929 (1997)
59. P M Coutinho, E Deleury, G J Davies, B Henrissat *J. Mol. Biol.* **328** 307 (2003)
60. C T Porter, G J Bartlett, J M Thornton *Nucleic Acids Res.* **32** 129 (2004)
61. G J Bartlett, C T Porter, N Borkakoti, J M Thornton *J. Mol. Biol.* **324** 105 (2002)
62. S F Altschul, W Gish, W Miller, E W Myers, D J Lipman *J. Mol. Biol.* **215** 403 (1990)
63. S Jones, J M Thornton *Curr. Opin. Chem. Biol.* **8** 3 (2004)
64. S Schmitt, D Kuhn, G Klebe *J. Mol. Biol.* **323** 387 (2002)
65. R A Laskowski, N M Luscombe, M B Swindells, J M Thornton *Protein Sci.* **5** 2438 (1996)
66. A V Tendulkar, P P Wangikar, M A Sohoni, V V Samant, C Y Mone *J. Mol. Biol.* **334** 157 (2003)
67. P J Artymiuk, A R Poirrette, H M Grindley, D W Rice, P Willett *J. Mol. Biol.* **243** 327 (1994)
68. R V Spriggs, P J Artymiuk, P Willett *J. Chem. Inf. Comput. Sci.* **43** 412 (2003)
69. A C Wallace, R A Laskowski, J M Thornton *Protein Sci.* **5** 1001 (1996)
70. A C Wallace, N Borkakoti, J M Thornton *Protein Sci.* **6** 2308 (1997)
71. J A Barker, J M Thornton *Bioinformatics* **19** 1644 (2003)
72. H Yao, D M Kristensen, I Mihalek, M E Sowa, C Shaw, M Kimmel, L Kavraki, O Lichtarge *J. Mol. Biol.* **326** 255 (2003)
73. C D Varfolomeev, K G Gurevich *Izv. Akad. Nauk, Ser. Khim.* 1629 (2001) [a]
74. C D Varfolomeev, I V Uporov, E V Fedorov *Biokhimiya* **67** 1328 (2002) [b]
75. A M Yaglom, I M Yaglom *Veroyatnost' i Informatsiya* (Probability and Information) (Moscow: Nauka, 1973)
76. C E Shannon, W Weaver *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois Press, 1949)
77. A G Wilson *Entropy in Urban and Regional Modelling* (London: Pion Ltd., 1970)
78. G R Lainen *Biokhimiya* **67** 107 (2002) [b]
79. E Di Cera, Q D Dang, Y M Ayala *Cell Mol. Life Sci.* **53** 701 (1997)
80. J B Cooper *Curr. Drug Targets* **3** 155 (2002)
81. C M Troy, M L Shelanski *Cell Death Differ.* **10** 101 (2003)
82. M Donepudi, M G Grutter *Biophys. Chem.* **101** 145 (2002)
83. Y Shi *Mol. Cell* **9** 459 (2002)
84. K Kolmodin, J Aqvist *FEBS Lett.* **498** 208 (2001)
85. D Barford, A K Das, M P Egloff *Annu. Rev. Biophys. Biomol. Struct.* **27** 133 (1998)
86. Z Y Zhang *Crit. Rev. Biochem. Mol. Biol.* **33** 1 (1998)
87. Z Jia Biochem *Cell Biol.* **75** 17 (1997)
88. J E Villafranca, C R Kissinger, H E Parge *Curr. Opin. Biotechnol.* **7** 397 (1996)
89. R J Williams *Mol. Biotechnol.* **23** 225 (2003)
90. B S Chevalier, B L Stoddard *Nucleic Acids Res.* **29** 3757 (2001)
91. M Galleni, J Lamotte-Brasseur, X Raquet, A Dubus, D Monnaie, J R Knox, J M Frere *Biochem. Pharmacol.* **49** 1171 (1995)
92. Z Wang, W Fast, A M Valentine, S J Benkovic *Curr. Opin. Chem. Biol.* **3** 614 (1999)
93. G Jung, H Ueno, R Hayashi *J. Biochem. (Tokyo)* **126** 1 (1999)
94. O Hlavacek, L Vachova *Folia Microbiol. (Praha)* **47** 203 (2002)
95. M Schmidt, A N Lupas, D Finley *Curr. Opin. Chem. Biol.* **3** 584 (1999)
96. V Tougu, T Kesvatera *Biochim. Biophys. Acta* **1298** 12 (1996)
97. A John, G Tuszynski *Pathol. Oncol. Res.* **7** 14 (2001)
98. S D Varfolomeev, A E Pozhitkov *Vestn. Mosk. Univ., Ser. 2: Khim.* **3** 147 (1999) [c]
99. B M Dunn *Chem. Rev.* **102** 4431 (2002)
100. J S Fruton *Q. Rev. Biol.* **77** 127 (2002)
101. N S Andreeva, A E Gustcina, A A Fedorov, N E Scutzkever, T N Volnova *Adv. Exp. Med. Biol.* **95** 23 (1977)
102. N S Andreeva, A S Zhdanov, A A Fedorov *Mol. Biol.* **18** 313b (1984) [d]
103. A R Sielecki, A A Fedorov, A Boodhoo *J. Mol. Biol.* **214** 148 (1990)
104. C L DiIanni, L J Davis, M K Holloway, W K Herber, P L Darke, N E Kohl, R A Dixon *J. Biol. Chem.* **265** 17348 (1990)
105. M Schulein *Biochim. Biophys. Acta* **1543** 239 (2000)
106. A Vasella, G J Davies, M Bohm *Curr. Opin. Chem. Biol.* **6** 619 (2002)
107. N C Strynadka, M N James *EXS* **75** 185 (1996)
108. B Fischer *EXS* **75** 143 (1996)
109. M Karplus, C B Post *EXS* **75** 111 (1996)
110. T Fukamizo *Curr. Protein Pept. Sci.* **1** 105 (2000)
111. J D Robertus, A F Monzingo *EXS* **87** 125 (1999)
112. I S Ridder, H J Rozeboom, K H Kalk, D B Janssen, B W Dijkstra *J. Biol. Chem.* **272** 33015 (1997)
113. K H Verschueren, F Seljee, H J Rozeboom, K H Kalk, B W Dijkstra *Nature (London)* **363** 693 (1993)
114. D G Knorre, S D Myzina *Biologicheskaya Khimiya* (Biological Chemistry) (Moscow: Vysshaya Shkola, 1992)
115. T M Glennon, Warshel *J. Am. Chem. Soc.* **120** 10 234 (1998)
116. D J Quirik, R T Raines *Biophys. J.* **76** 1571 (1991)
117. B C Oh, W C Choi, S Park, Y O Kim, T K Oh *Appl. Microbiol. Biotechnol.* **63** 362 (2004)
118. J W Kozarich *Curr. Opin. Chem. Biol.* **7** 78 (2003)
119. M Holmquist *Curr. Protein. Pept. Sci.* **1** 209 (2000)
120. F Drablos, S B Peterson, in *Methods in Enzymology* Vol. 284 (Eds B Rubin, E A Dennins) (San Diego, CA: Academic Press, 1997) p. 28
121. S D Varfolomeev, K G Gurevich *Biokinetika* (Biokinetics) (Moscow: Fair-Press, 1992)
122. K Martinek, V N Dorovska, S D Varfolomeev, I V Berezin *Biochim. Biophys. Acta* **271** 8086 (1972)
123. B H Rehm *Biochem. J.* **376** 15 (2003)
124. P Wooley, S Petersen (Eds) *Lipases, Their Structure, Biochemistry and Applications* (Cambridge: Cambridge University Press, 1994)
125. H Soreq, A Gnatt, Y Loewenstein, L F Neville *Trends Biochem. Sci.* **17** 353 (1992)
126. G Dodson, A Wlodamer *Trends Biochem. Sci.* **23** 347 (1998)
127. E B Fauman, M A Saper *Trends Biochem. Sci.* **21** 413 (1996)
128. M Paetzel, R E Dalbey *Trends Biochem. Sci.* **22** 28 (1997)
129. C Goffin, J M Ghuysen *Microbiol. Mol. Biol. Rev.* **62** 1079 (1998)
130. G J Palm, J Lubkowski, C Derst, S Schleper, K H Rohm, A Wlodawer *FEBS Lett.* **390** 211 (1996)
131. S J Yoo, Y K Shim, I S Seong, J H Seol, M S Kang, C H Chung *FEBS Lett.* **412** 57 (1997)
132. H J Duggleby, S P Tolley, C P Hill, E J Dodson, G Dodson, P C Moody *Nature (London)* **373** 264 (1995)
133. G L Eichhorn (Ed.) *Inorganic Biochemistry* (Amsterdam, Oxford, New York: Elsevier, 1973)
134. E Kimura *Curr. Opin. Chem. Biol.* **4** 207 (2000)
135. C Montecucco, G Schiavo *Trends Biochem. Sci.* **18** 324 (1993)
136. K M Holtz, E R Kantrowitz *FEBS Lett.* **462** 7 (1999)
137. J E Murphy, B Stec, L Ma, E R Kantrowitz *Nature Struct. Biol.* **4** 618 (1997)
138. K M Holtz, E R Kantrowitz *FEBS Lett.* **462** 7 (1999)
139. S Benini, W R Rypniewski, K S Wilson, S Miletti, S Ciurli, S Mangani *Struct. Fold Des.* **7** 205 (1999)
140. I G Danilova, A D Ryabov, S D Varfolomeev *J. Mol. Catal.* **118** 161 (1997)
141. V S Sergeeva, E N Efremenko, G M Kazankov, A K Gladilin, S D Varfolomeev *Biotechnol. Tech.* **13** 479 (1999)

142. J K Crimsky, J M Scholtz, C N Pace, J R Wild *Biochemistry* **36** 14 366 (1992) [b]
143. K Lay, K I Dave, J R Wild *J. Biol. Chem.* **269** 16579 (1994)
144. G A Omburo, J M Kuo, L S Mullins, F M Rushel *J. Biol. Chem.* **267** 13278 (1992)
145. V S Sergeeva, E N Efremenko, G M Kazankov, S D Varfolomeev *J. Mol. Catal. B: Enzym.* **10** 571 (2000)
146. S M Avaeva *Biokhimiya* **65** 441 (2000) [b]
147. G S Baldwin, N A Gormley, S E Halford *Met. Ions Biol. Syst.* **37** 345 (2000)
148. P H Burton, D C Hall, J Josse *J. Biol. Chem.* **245** 4346 (1970)
149. S M Avaeva, T I Nazarova *Usp. Biol. Khim.* **26** 42 (1985)
150. V Yu Oganessyan, S A Kurilova, N N Vorobyova, T I Nazarova, A N Popov, A A Lebedev, S M Avaeva, E H Harutyunyan *FEBS Lett.* **348** 301 (1974)
151. E G Artyunyan, V Yu Oganesyan, N N Oganesyan, S S Terzyan, A N Popov, S V Rubinskii, B K Vainshtein, T I Nazarova, S A Kurilova, N N Vorob'eva, S M Avaeva *Kristallografiya* **41** 84 (1996) [e]
152. E H Harutyunyan, V Yu Oganessyan, N N Oganessyan, S M Avaeva, T I Nazarova, N N Vorobyova, S A Kurilova, R Huler *Biochemistry* **35** 7754 (1996) [b]
153. P Heikinheimo, V Tuominen, A K Ahonen, A Teplyakov, B S Cooperman, A A Baykov, R Lahti, A Goldman *Proc. Natl. Acad. Sci. U.S.A.* **98** 3121 (2001)
154. G Lukatela, N Krauss, K Theis, T Selmer, V Gieselmann, K von Figura, W Saenger *Biochemistry* **37** 3654 (1998) [b]
155. T A Steitz *J. Biol. Chem.* **274** 17395 (1999)
156. D Sam, G G Perova *Biochemistry* **38** 6576 (1999) [b]
157. G S Baldwin, R B Sessions, S G Erskine, S E Galford *J. Mol. Biol.* **288** 87 (1999)
158. H Viadiu, A K Aggarwal *Nature Struct. Biol.* **5** 910 (1998)
159. L Holm, C Sander *Science* **273** 595 (1996)
160. I G Gazaryan *Biotekhnologiya (Itogi Nauki i Tekhniki)* [Biotechnology (Advances in Science and Engineering Series)] (Moscow: Izd. VINITI, 1992) Vol. 34, p. 4
161. B C Finzel, T L Poulos, J Kraut *J. Biol. Chem.* **259** 13027 (1984)
162. N Kunishima, K Fukuyama, H Matsubara, H Hatanaka, Y Shibano, T Amachi *J. Mol. Biol.* **235** 331 (1994)
163. J J Birtoft, L J Banaszak *J. Biol. Chem.* **258** 472 (1983)
164. A R Clarke, H M Wiks, D A Barstow, T Atkinson, N N Chia, J J Holbrook *Biochemistry* **27** 1617 (1988) [b]
165. H Taguchi, T Ohta, H Matsuzawa *J. Biochem. (Tokyo)* **122** 802 (1997)
166. E A MacGregor, S Janecek, B Svensson *Biochim. Biophys. Acta* **1546** 1 (2001)
167. M L Rabinovich, M S Mel'nik, A V Bolobova *Biokhimiya* **67** 1026 (2002) [b]
168. P Tomme, R A Warren, N R Gilkes *Adv. Microb. Physiol.* **37** 1 (1995)
169. P Beguin *Annu. Rev. Microbiol.* **44** 219 (1990)
170. P Beguin, P M Alzari *Biochem. Soc. Trans.* **26** 178 (1998)
171. N R Gilkes, B Henrissat, D G Kilburn, R C Miller, R A Warren *Microbiol. Rev.* **55** 303 (1991)
172. M Hassouni, B Henrissat, M Chippaux, F Barras *J. Bacteriol.* **174** 765 (1992)
173. G A Tavares, P Beguin, P M Alzari *J. Mol. Biol.* **273** 701 (1997)
174. M E Himmel, W S Adney, J W Fox, D J Mitchell, J O Baker *Appl. Biochem. Biotechnol.* **39–40** 213 (1993)
175. V I Krupyanko *Biokhimiya* **53** 2059 (1988) [b]
176. M V Nogues, M Vilanova, C M Cuchillo *Biochim. Biophys. Acta* **1253** 16 (1995)
177. D M Blow *Trends Biochem. Sci.* **22** 405 (1997)
178. S Bencharit, C L Morton, Y Xue, P M Potter, M R Redinbo *Nature Struct. Biol.* **10** 349 (2003)
179. Y S Ho, L Swenson, U Derewenda, L Serre, Y Wei, Z Dauter, M Hattori, T Adachi, J Aoki, H Arai, K Inoue, Z S Derewenda *Nature (London)* **385** 89 (1997)
180. C G Swain, J I Brown *J. Am. Chem. Soc.* **74** 2534 (1952)
181. C G Swain, J I Brown *J. Am. Chem. Soc.* **74** 2538 (1952)
182. K B Schowen, H H Limbach, G S Denisov, R L Schowen *Biochim. Biophys. Acta* **1458** 43 (2000)
183. M Eigen *Angew. Chem., Int. Ed. Engl.* **3** 1 (1964)
184. J H Wang *Science* **161** 328 (1968)
185. N G Esipova, V G Tumanyan *Mol. Biol.* **6** 840 (1972) [d]
186. M Sasai *Proc. Natl. Acad. Sci. U.S.A.* **92** 8438 (1995)
187. I Friedberg, T Kaplan, H Margalit, in *Proceedings of International Conference on Intelligent Systems for Molecular Biology, San Diego, CA, 2000* Vol. 8, p. 162
188. Z Q Cheng, B A McFadden *Protein Eng.* **11** 457 (1998)
189. A C Martin, J M Thornton *J. Mol. Biol.* **263** 800 (1996)
190. A V Efimov *FEBS Lett.* **355** 213 (1999)
191. S D Lahiri, G Zhang, J Dai, D Dunaway-Mariano, K N Allen *Biochemistry* **43** 2812 (2004) [b]
192. M Saito, K Hori, T Kurotsu, M Kanda, Y Saito *J. Biochem. (Tokyo)* **117** 276 (1995)
193. K E Jaeger, B W Dijkstra, M T Reetz *Annu. Rev. Microbiol.* **53** 315 (1999)
194. M Eigen, L De Mayer *Z. Electrochem.* **59** 986 (1955)
195. M Eigen *Discuss. Faraday Soc.* **39** 7 (1965)

[a] — *Russ. Chem. Bull., Int. Ed. (Engl. Transl.)*
[b] — *Biochemistry (Moscow) (Engl. Transl.)*
[c] — *Moscow Univ. Bull. (Engl. Transl.)*
[d] — *Mol. Biol. (Engl. Transl.)*
[e] — *Crystallogr. Rep. (Engl. Transl.)*