Active Sequences Collection (ASC) database: a new tool to assign functions to protein sequences

Angelo M. Facchiano*, Antonio Facchiano1 and Francesco Facchiano1

Istituto di Scienze dell’Alimentazione, CNR—via Roma 52A/C 83100 Avellino, Italy and 1IDI, Istituto Dermopatico dell’Immacolata—via Monti di Creta 104, 00167 Roma, Italy

Received August 14, 2002; Revised and Accepted September 25, 2002

ABSTRACT

Active Sequences Collection (ASC) is a collection of amino acid sequences, with an unique feature: only short sequences are collected, with a demonstrated biological activity. The current version of ASC consists of three sections: DORRS, a collection of active RGD-containing peptides; TRANSIT, a collection of protein regions active as substrates of transglutaminase enzyme (TGase), and BAC, a collection of short peptides with demonstrated biological activity. Literature references for each entry are reported, as well as cross references to other databases, when available. The current version of ASC includes more than 800 different entries. The main scope of this collection is to offer a new tool to investigate the structural features of protein active sites, additionally to similarity searches against large protein databases or searching for known functional patterns. ASC database is available at the web address http://crisceb.unina2.it/ASC/ which also offers a dedicated query interface to compare user-defined protein sequences with the database, as well as an updating interface to allow contribution of new referenced active sequences.

INTRODUCTION

Sequencing the genome of humans, as well as of other organisms, led to hundreds of thousands of DNA sequences for which the corresponding predicted proteins have unassigned functions. Understanding the genome requires a multidisciplinary approach including prediction of functions from the sequence. Bioinformatics tools need to be continuously updated and new strategies are necessary to achieve this challenging aim. Commonly used protein sequence databases such as SWISS-PROT, TrEMBL, PIR, GENPEPT, are mainly aimed at collecting the largest possible number of known sequences, more- or less-detailed annotations about their biological activity, post-translational modifications, active sites or structural/functional domains. These databases are commonly used to find structural similarities, and eventually to hypothesize the function of newly identified proteins. Such databases may be considered ‘not function-oriented’, as compared to ‘function-oriented’ databases like PROSITE, PRINTS, PRODOM, which collect sequence information related to specific activities, and other activity-specialized databases as MHC-binding peptides or HIV epitopes collections. Despite the availability of such databases and their related searching tools, functional sites within proteins are often difficult to be predicted by bioinformatics tools devoted to the analysis of amino acid sequences. In many cases, functional sites consist of a few amino acids in a small pocket, not contiguous in the sequence but very close in their three-dimensional arrangement; in these protein families a simple sequence pattern characterizing such functional site may not exist. In other cases, the sequence around the functional amino acids may be more relevant, but any sequence pattern might be nonspecific or redundant or inadequate. Further, within the whole sequence of a protein, functional sites represent a very small part of the whole, being hidden within ‘not-reactive’ regions forming the protein scaffold. Sometimes, proteins may contain more than one functional site, such as moonlight proteins (1), and this confers an even higher level of complexity to the prediction. Therefore, active sequences are often hidden and in many cases difficult to identify.

In this work, an Active Sequences Collection (ASC) has been created aiming at developing new tools to help assign biological functions to a given protein under investigation. This unique collection includes only short bioactive regions of protein sequences, making it possible to compare a given entire protein sequence with the collected protein fragments showing known functions.

ASC consist, at the present time, of three sections: DORRS (Database Of RGD-Related Sequences), containing RGD-related molecules with known function; TRANSIT (TRANsglutamination SITes), a collection of protein regions with demonstrated activity as transglutaminase substrates, and BAC (BioACtive peptides), a collection of peptides active

*To whom correspondence should be addressed. Email: angelo.facchiano@isa.av.cnr.it

The authors wish it to be known that, in their opinion, all three authors should be regarded as joint First Authors
in vitro or in vivo. Each entry reports literature references and annotations, and cross references to other databases when available. The entire ASC collection consists currently of more than 800 entries and is periodically updated by a systematic review of the upcoming literature. Moreover, any researcher can contribute active, fully referenced sequences to the collection. New sections are planned to be available shortly.

ASC database is accessible at the web address http://crisceb.unina2.it/ASC/ where the user finds the general description, references and specific tools to browse the collections. ASC consists of text files indexed by the SRS (Sequence Retrieval System), available at the same web site. Moreover, a novel PERL program enables searching a given amino acid sequence against the ASC databases and to verify any found similarity. This search may represent an useful tool to investigate protein sequences and to find structural similarities with short sequences with known biological activity.

DORRS

Molecules containing the short RGD (Argine–Glycine–Aspartic acid) motif are largely investigated for their anti-adhesive activity (2–5). Such function makes these molecules interesting candidates as anti-metastasis drugs currently under clinical investigation (6–10). This motif is present in many proteins including components of the extracellular matrix; by interacting with integrins, it mediates cell-matrix as well as cell–cell interaction. It also represents a key active site of disintegrins, potent anti-aggregant molecules found in snake venoms (11). Recently, RGD-containing peptides have been shown to induce apoptosis, with an integrin-independent mechanism. Such direct apoptotic effect has been demonstrated on normal as well as tumoral cells (12–15). RGD-containing peptides released from the matrix have been suggested to play a role in the bone remodeling process (16); the RGD motif is also studied in gene therapy as a delivery-system, by exploiting its ability to specifically target integrins (17,18). Finally, RGD-related peptides have been shown to play a key role in the modulation of the immune response (19,20).

While the RGD motif is a rather nonspecific site, the surrounding residues at the N and C terminus are known to contribute specificity. Active linear and cyclized short peptide sequences containing RGD were extracted from the literature, as well as molecules mimicking RGD-peptide, for a total of 113 molecules (July 2002 release). Bibliographic references were also collected for each sequence. Further, physical-chemical properties as molecular weight and isoelectric point (IP) were calculated and reported, for the linear sequences. When available, kinetic parameters as Kd or IC₅₀ values were also reported. A subset containing molecules showing a partial activity was also created, according to the definition reported in the corresponding bibliographic reference. An additional feature is the listing of referenced non-active sequences. This represents a novel feature; in fact, besides the active-sequences, non-active sequences may be useful to understand crucial properties for gain or loss of activity, and in designing novel active molecules. Further, this feature may provide useful control peptides to assay a specific activity. To our knowledge, this is the first available database of RGD-related sequences grouped as active and non-active molecules and reporting physical-chemical properties.

TRANSIT

Transglutaminase (TGase, E.C. 2.3.2.13) is an ubiquitous class of enzymes whose functions are largely investigated. In fact, a de-regulation of TGase’s functions has been shown to be related to a number of human pathologies like coeliac disease, coagulation disorders, cancer, neurodegenerative diseases and others. Despite the large interest about such enzyme, TGases mechanisms of action in both physiologic and pathologic conditions are poorly identified and still under investigation (21–23). For instance, the role of TGase type II, or tissue TGase, as a pro-apoptotic player is a fascinating and intriguing field of interest. In fact, a large number of papers showed that in many apoptotic models tissue TGase is potently activated, leading to the formation of covalent bonds among different intracellular or extracellular proteins (24–28). More recently, the effect of tissue TGase on apoptosis has been shown to be highly dependent on the type of the apoptotic stimuli and the way crosslinking activity is affected (28,29). On the other hand, tissue TGase has also been recently shown to provide a protection against apoptotic insults (30), suggesting that the role of TGase in apoptotic processes is still not completely elucidated. The identification of the protein substrates, cross-linked by TGase under physiological and pathological conditions, is a very hot but difficult topic. In fact, high molecular weight complexes formed by the action of this enzyme are often very difficult to separate and analyze, even with the modern proteomics approaches.

Members of this enzyme-family are highly homologous and have been shown to exert different actions (crosslinking, G-protein and GTPase, ATPase, deamidase activity and others), although the structural features underlying such functional differences are not yet well known. The crosslinking activity consist of a transamidation reaction forming a covalent bond between the amide group of a glutamine side chain and the amino group of an amine donor, i.e. a polyamine or a lysine side chain, with ammonia release. The specificity of the amino acids surrounding the glutamine and the lysine residues involved in transglutamination is still debated and under investigation, as it was pointed out recently by researchers at the WHAT web site (http://crisceb.unina2.it/what/) devoted to discussions and information around this enzyme family. The problem is rather complex, since some members of TGase family are present and active in the extracellular environment, while others are inside the cell and can be found in the cytoplasm, vesicular compartments, or both, but can also be secreted outside the cell. Consequently, transglutamination may occur in environments very different for ionic strength, substrate accessibility, hydrophobicity, calcium and nucleotide concentration. It is noteworthy that while some TGase isoforms are strictly calcium-dependent, others are modulated by nucleotides too. Therefore, not surprisingly, previous studies aimed at identifying a structural pattern of amino acids surrounding the reactive glutamine gave contrasting results. While it is evident that only specific glutamine residues in proteins may act as acyl donors,
the specific sequence pattern identifying glutamine as substrate is not known. A computational approach may significantly help in this case: in fact, a given protein sequence can be compared with protein-regions containing glutamines or lysines shown to be TGase substrates. Such regions have been collected in TRANSIT. The current release (July 2002) consists of 63 entries, each containing the literature reference and annotation regions clustered in subgroups for a specified TGase isofrom. This database can be accessed via the web interface, to evaluate the similarity of the sequence environment surrounding the glutamine or lysine within a given protein with known substrates of TGases. Hence, TRANSIT may help identify reactive residues as putative TGase substrate. In the near future, it is planned to improve the information on specific TGase isofroms.

**BAC**

This is a collection of biologically active peptides derived from the literature. In contrast with the other two ASC sections, BAC is not oriented to a specific protein function, being aimed to a more general investigation of sequence-function relationships. It contains more than 650 entries (in July 2002 release) and represents the largest collection of short active sequences freely available on the web and a powerful tool to search a given protein sequence for similarities with peptides known to exhibit biological activity. Similarity searches carried out on BAC will escape the problems related to the redundancy and the ‘noise’ experienced on the larger protein sequences databases, in which a large part of the collected information is either redundant or not relevant to investigate short active regions. Searching BAC can be helpful when an active region is to be identified within a whole protein sequence or when an active peptide has to be designed; in the latter case, sequences sharing homology with known active peptides, as well as negative control peptides sharing no homology with any other peptide, are sought.

It should be evidenced that BAC does not collect sequence patterns and signature sequences; rather, it includes only sequences shown to have full referenced biological activity as peptides.

**FUTURE DIRECTIONS**

ASC is aimed to create a bioinformatics resource for scientists interested to investigate structure-function relationships of proteins and peptides. On this basis, we plan to expand the collection by increasing the number of entries of the existing sections, as well as by creating new sections oriented to specific functions. As an example, a new section is under construction, devoted to peptides with relevant interest in food science. Moreover, additional information will be added to the existing entries, by creating new fields as ‘keywords’ and improving the linking to other databases. An interactive form is available on the web, and the scientific community is invited to contribute information suitable to be added in ASC. Any submission of new entry or improvement to existing entries will be accepted, provided that the information to be added in ASC are fully referenced.

**REFERENCES**


