# Construction of A Non-Redundant Human SH2 Domain Database

Haiming Huang, Yuchen Jiao, Rui Xu, and Youhe Gao*

*Department of Pathophysiology/National Key Laboratory of Medical Molecular Biology/Proteomics Research Center, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing 100005, China.*

**Domain database is essential for domain property research. Eliminating redundant information in database query is very important for database quality. Here we report the manual construction of a non-redundant human SH2 domain database. There are 119 human SH2 domains in 110 SH2-containing proteins. Human SH2s were aligned with ClustalX, and a homologous tree was generated. In this tree, proteins with similar known function were classified into the same group. Some proteins in the same group have been reported to have similar binding motifs experimentally. The tree might provide clues about possible functions of hypothetical proteins for further experimental verification.**

**Key words: SH2 domain, non-redundant database, homologous tree**

## Introduction

Since the start of the Human Genome Project, the public databases have been growing rapidly. These explosively increasing information revolutionized the biology research. However, there are too many redundant data confusing researchers. For example, when we search the Genbank for the human Nck1 protein, we receive six different protein entries. But they all have the same amino acid sequence and denote the same protein, human Nck1. The difference is mainly on the description of the protein name, for example, NCK adaptor protein 1, Cytoplasmic protein NCK1, nck protein-human, unnamed protein product, and so on.

The importance of modular proteins in biology and human diseases is emphasized by the recent observation that the majority of positionally cloned human disease genes encode multidomain proteins, many of which are, in fact, signaling proteins (*1*). The SH2 domains (Src homology 2) serve as the prototype for a growing family of protein-interaction modules; its polypeptides are involved in transmitting signals from external and internal cues (*2*). This globular domain of approximately 100 amino acids has a pocket that directly binds the phosphotyrosine moiety of phosphoproteins or phosphopeptides (*3*). Characterization of the human SH2 protein will help us to understand the secret of cellular signaling and disease therapy. To study the properties of human SH2s, it is necessary to build a non-redundant human SH2 domain database besides a protein database containing the SH2 domains. Currently, the commonly used tools for domain query are CDART (Conserved Domain Architecture Retrieval Tool; ref. *4*) in NCBI and SMART (Simple Modular Architecture Research Tool; ref. *5*), by which many SH2-containing proteins can be found. However, the results are usually redundant. A complete *non-redundant* human SH2 *domain* database has not been found yet with our best effort. We believe that human inspection is required to make a high-quality non-redundant domain database. In this report, based on CDART and SMART search results, we manually constructed a non-redundant human SH2 domain database. With multi-alignment program ClustalX, the SH2 domains were aligned and a homologous tree was generated, both of which may provide clues for experimental study of SH2 domain functions.

*** Corresponding author.**
**E-mail: gaoyouhe@pumc.edu.cn**

## Results and Discussion

### Construction of a non-redundant human SH2 database

CDART is a search tool to perform similarity searches of the NCBI Entrez Protein Database based on do-

main architecture, defined as the sequential order of conserved domains in proteins, while SMART allows rapid identification and annotation of signaling domain sequences. By these methods, 200 and 196 human SH2 protein sequences were obtained from NCBI Entrez Protein Database and SMART, respectively. In these 396 sequences, some are the same SH2 protein sequences with different description; some are the protein fragments of full-length proteins. The SH2 domain range of each SH2 protein was firstly determined by Motif Scan (http://hits.isb-sib.ch/cgi-bin/PFSCA). Then, all of the redundant SH2 domains were eliminated as described in the materials and methods. As a result, a non-redundant human SH2 domain database with 110 unique sequences of SH2-containing proteins was constructed. Because some SH2 proteins, for example phospholipase C gamma 1 and gamma 2, have two SH2 domains, there are totally 119 different SH2 domains in the database. However, our non-redundant SH2 database should be updated as changes of CDART and SMART occur. This database is available from http://www.proteomics-cams.com/service/database-sh2.htm.

### Multiple alignments

These 119 different SH2 domain sequences were aligned with ClustalX (1.8) and a homologous tree was built (Figure 1). The proteins from one family were clustered into one group, such as STATs, Tensins, JAKs, SOCSs, VAVs, GRBs, chimerins and SHPs families, which is consistent with published results. Some proteins in one group were found to have the same or similar binding motifs according to published data. For example, the proteins FYN and v-fgr share the same binding motif YEEI (*3*) and have a sequence identity of 83% (Figure 2A), which endows them similar function and binding motif. Another example is SH2 domain protein 1A (SH2D1A) and EAT-2, which also have similar binding pattern, with the former has a binding motif of YXXV/I (X denotes any amino acid) and the latter has a binding motif of YAQV (*6*), although their sequence identity of 43.93% is relatively low (Figure 2B).

Some hypothetical proteins are grouped with known proteins, such as hypothetical protein FLJ11700 and ras inhibitor, hypothetical protein FLJ00138 and SHB, hypothetical protein FLJ14886 and SH2 domain protein 2A (SH2D2A). Their sequence identities are 38.39%, 56.76%, and 36.94%, respectively (Figure 3). Based on the homologous tree we built, it suggests that some hypothetical proteins have the similar binding motifs and functions to their known similar proteins.

Non-redundant domain databases are indispensable for functional study of these domains. Here, we manually constructed a non-redundant human SH2 domain database containing 119 unique SH2 domains. To our knowledge, it has been the most complete non-redundant human SH2 domain database so far. We think that the finding of numbers of human SH2 domains, sequence relation of SH2 domains, and prediction of hypothetical SH2 domain function are useful information for SH2 domain researchers. We have used the information to construct a clone library of 80 human SH2 domains for studying their binding properties (*7*). Even though we agree that further experimental confirmations are absolutely required, we believe that this database provides useful information for domain property research and is an interesting clue for researchers.

## Materials and Methods

### Protein database containing human SH2 domains

The CDART Querying was used for searching the CDART website in the NCBI Genbank (http://www.ncbi.nlm.nih.gov/BLAST/) for all of the human SH2 proteins. The result with 200 entries was saved in a Microsoft Word file. The SMART Querying was used for searching the SMART website (http://smart.embl-heidelberg.de/) for all of the human SH2 proteins. The result with 196 entries was saved in another Microsoft Word file.

### Definition of the SH2 domain

The SH2 domain ranges of each SH2 protein were determined by Motif Scan in http://hits.isb-sib.ch/cgi-bin/PFSCAN.

### Elimination of redundant entries

The first SH2 domain from the CDART querying was put in a new Word file; the second SH2 domain was compared with the first one by the Find command of Microsoft Word for exact match. The same domains were excluded and the other were listed as the second entry and saved in the database file. A non-redundant

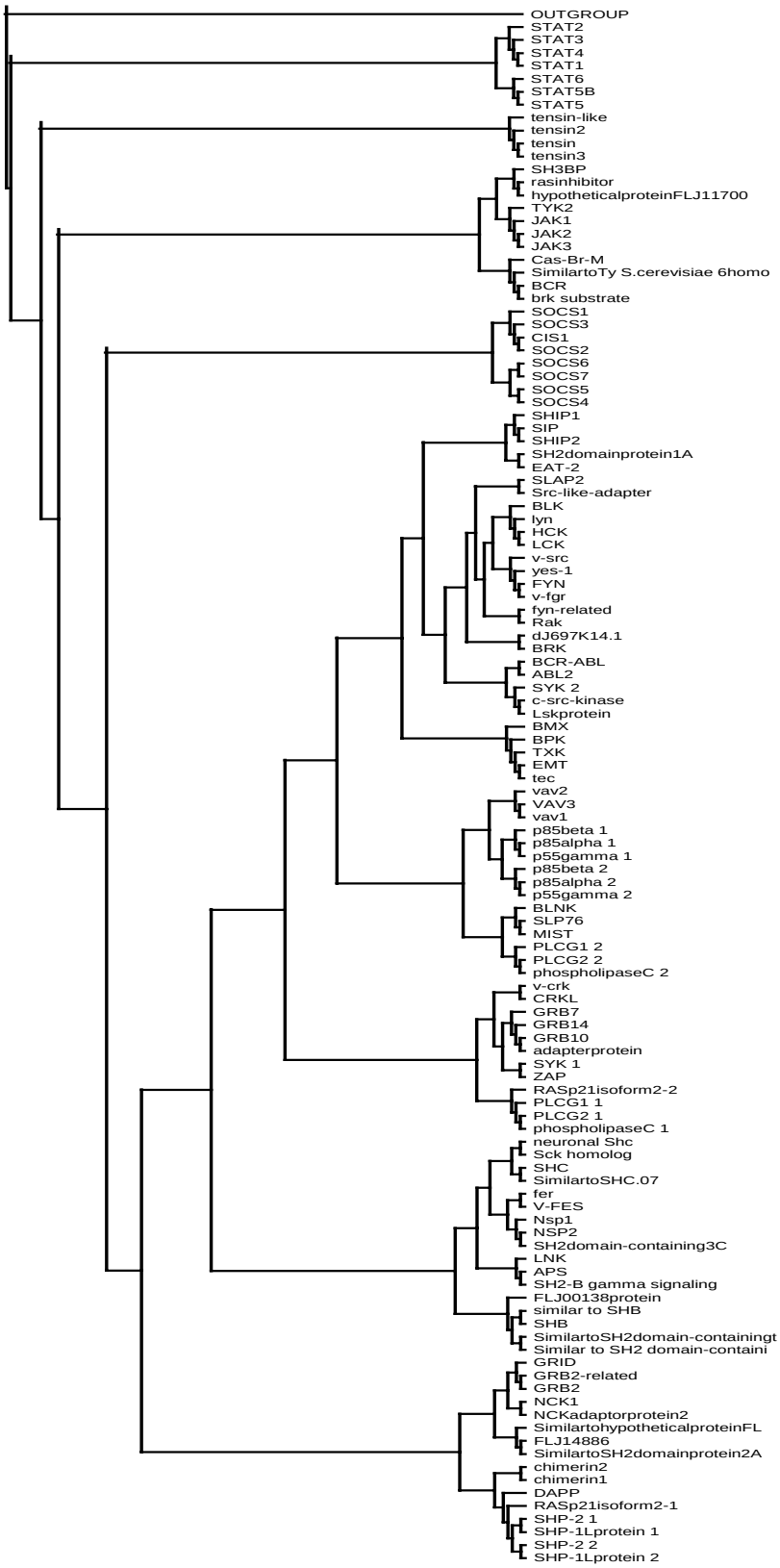


**Fig. 1** The homologous tree of all SH2 domains in the non-redundant database.

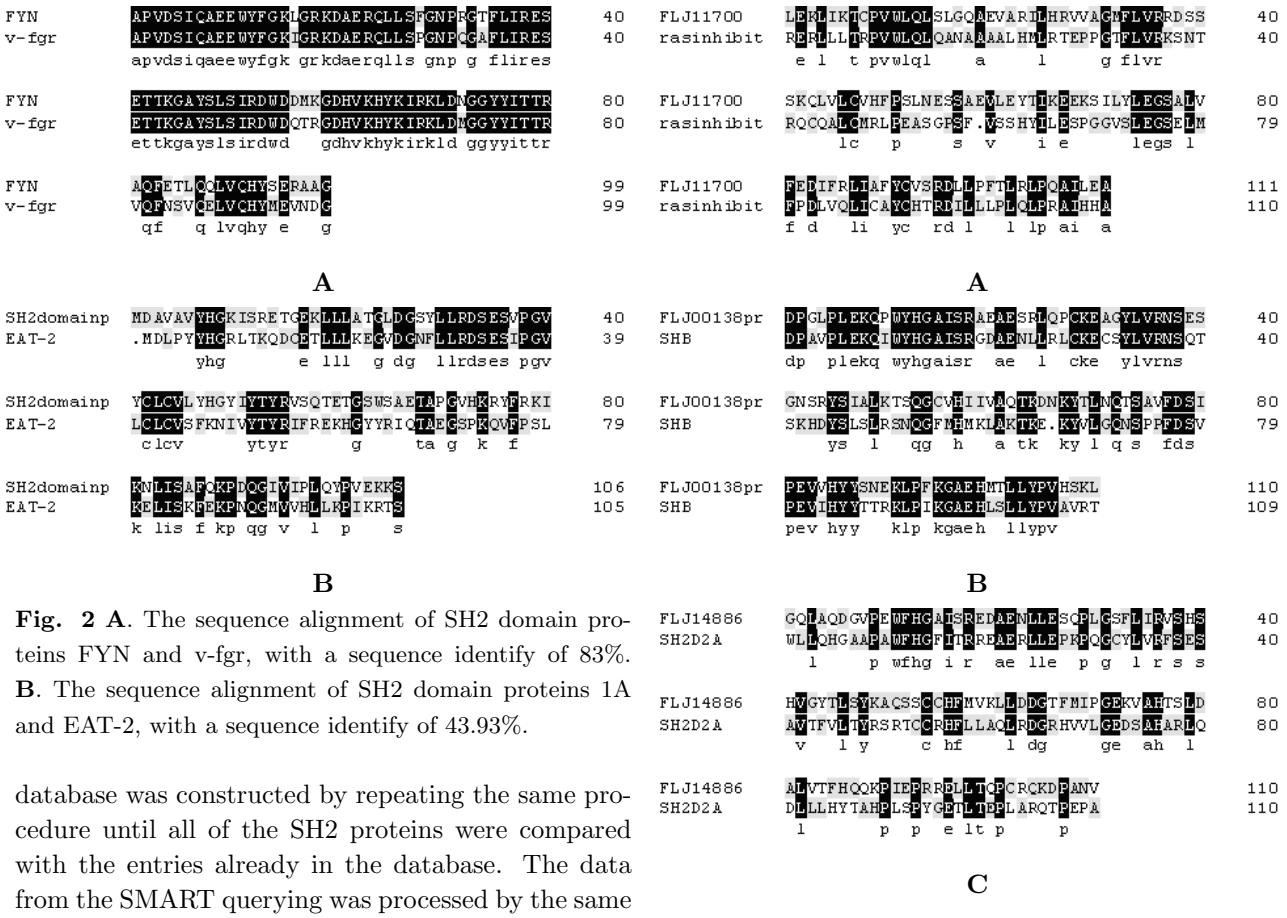


**Fig. 2 A**. The sequence alignment of SH2 domain proteins FYN and v-fgr, with a sequence identify of 83%. **B**. The sequence alignment of SH2 domain proteins 1A and EAT-2, with a sequence identify of 43.93%.

**Fig. 3** The sequence alignment of hypothetical protein FLJ11700 and ras inhibitor (A), hypothetical protein FLJ00138 and SHB (B), hypothetical protein FLJ14886 and SH2 domain protein 2A (SH2D2A) (C). The sequence identities of them are 38.39%, 56.76% and 36.94%, respectively.

database was constructed by repeating the same procedure until all of the SH2 proteins were compared with the entries already in the database. The data from the SMART querying was processed by the same procedure.

### Multiple Alignment

All the sequences of the non-redundant database were aligned by ClustalX (1.8) and a homologous tree was built.

## References


1. Mushegian, A.R., *et al.* 1997. Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc. Natl. Acad. Sci. USA* 94: 5831-5836.
2. Pawson, T., *et al.* 2001. SH2 domains, interaction modules and cellular wiring. *Trends Cell Biol.* 11: 504-511.
3. Songyang, Z. and Cantley, L.C. 1995. Recognition and specificity in protein tyrosine kinase-mediated signalling. *Trends Biochem. Sci.* 20: 470-475.
4. Geer, L.Y., *et al.* 2002. CDART: protein homology by domain architecture. *Genome Res.* 12: 1619-1623.
5. Schultz, J., *et al.* 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA* 95: 5857-5864.
6. Li, C., *et al.* 2003. Dual functional roles for the X-linked lymphoproliferative syndrome gene product SAP/SH2D1A in signaling through the signaling lymphocyte activation molecule (SLAM) family of immune receptors. *J. Biol. Chem.* 278: 3852-3859.
7. Ma, S., *et al.* 2003. Rapid method of constructing domain library. *Chin. J. Biochem. Mol. Biol.* 19: 537-541.


This work was partly supported by grants from National Natural Science Foundation of China (No. 3037030, 30270657 and 30230150), Major State Basic Research Development Program of China (2004CB520804), Pilot Study for Key Basic Research Project of China (2002CCA04100), and Key Project for International Cooperation of China (2002AA229031).