BIOCHEMICAL JOURNAL LETTERS

Evolution of the hyaluronan-binding module of link protein

Link protein (LP), which in cartilage stabilizes the interaction between aggrecan and hyaluronan, is assembled from three separate modules [1–4]. The 125-amino-acid-long module L1 shows a distant structural relationship with the variable region of immunoglobulin chains [3,5] and is involved in binding to cartilage proteoglycan [6]. Modules L2 and L3, the hyaluronan-binding modules (HABM), are tandem repeats of about 100 amino acid residues, and were shown to interact with hyaluronan [6,7]. These modules are encoded by separate exons and show homology to other multidomain proteins. All three module types have been found in the G1 domains of the aggregating proteo-glycans aggrecan [8–12], versican [13] and neurocan [14]. L2 and L3 are represented once more in the G2 domain of aggrecan. Recently, the cell-surface adhesion receptor CD44 [15,16] and a secreted tumour-necrosis-factor-inducible protein (TSG-6) [17] were shown to possess a single-copy HABM. To sum up, in a single genome there must be at least 12 genetic elements encoding

the HABMs, which have been generated by exon duplication and shuffling. In order to suggest the order of duplication events, we performed a detailed sequence comparison of 32 HABMs and 9 Ig-like modules by the PileUp program of the Genetic Computer Group, University of Wisconsin [18]. The scoring table of Gribkov and Burgess [19] was used with scores adjusted to a mean of 0 and standard deviation of 1.0.

On the basis of the similarity of all sequence pairs, the clustering relationships were plotted as dendrograms and multiple alignments were made. As shown for the HABMs (Figure 1), four cysteines, three glycines, two alanines and one proline are invariant. Furthermore, only aliphatic amino acids occupy positions 37, 54 and 59, and aromatic residues are found at positions 53 and 103. These amino acids must be essential for the stable folding of the module and cannot be replaced by others.

As shown in Figure 1, among the HABMs the most distant relative is CD44. Modules L2 seem to be more similar to each other than to L3 modules. Modules of link protein and proteo-glycans represent two separate subgroups. The closest relatives are the G1 and G2 domains of aggrecan, although the aggrecan
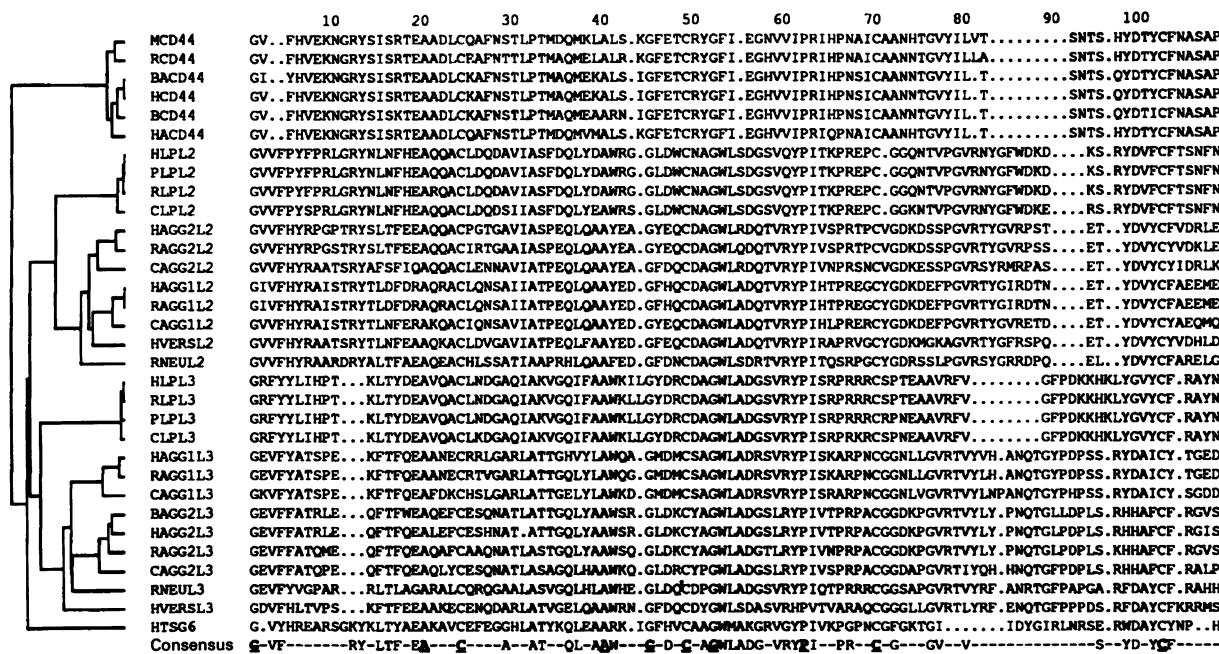
Figure 1   Dendrogram representing the clustering relationship and multiple alignment of the hyaluronan-binding modules

Sequences were compared for mouse (M; [20]), rat (R; [21]), baboon (BA; [22]), human (H; [15,16]), bovine (B; [23]) and hamster (HA; [24]) CD44, human [4], pig (P; [3]), rat [1] and chicken (C; [2]) link protein (LP), human [11], rat [9], chicken [12] and bovine [10] aggrecan (AGG), human versican (VERS) [13], rat neurocan (NEU) [14] and human TSG-6 [17]. Module borders match with exon borders in the link-protein genes [25,26]. Consensus residues are present in more than half of the sequences. Residues identical in all sequences are printed **bold** and underlined in the consensus line.
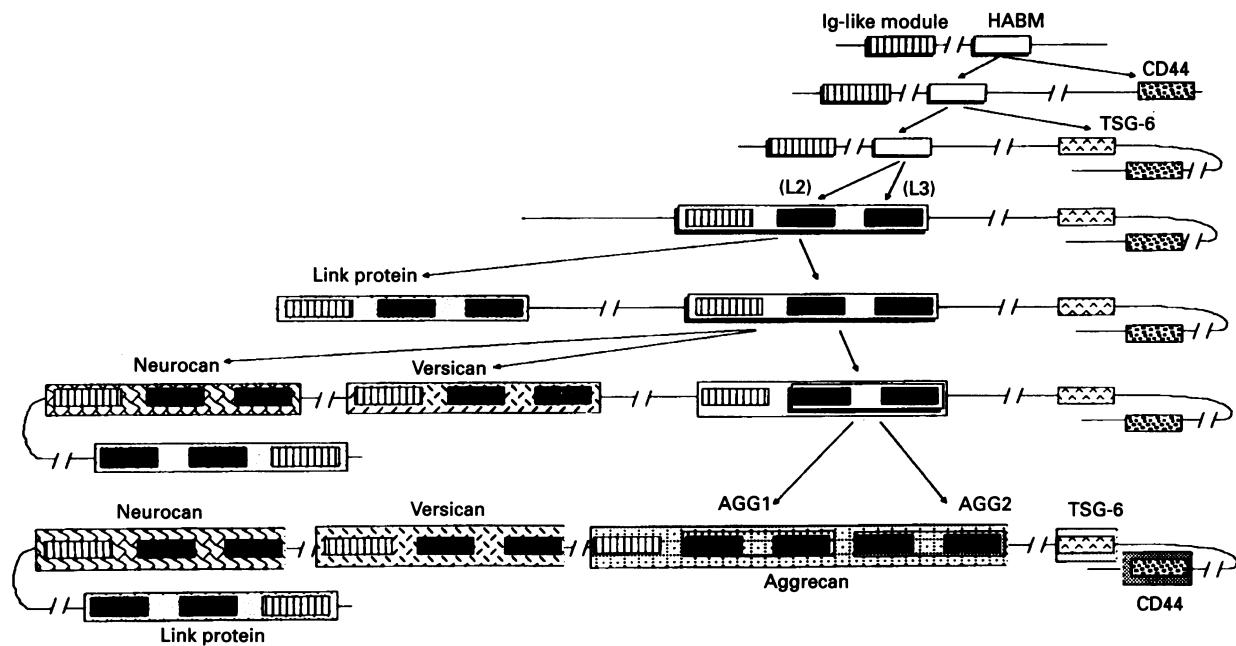
Figure 2   Suggested evolutionary scheme for the hyaluronan-binding module

Genetic elements for HABMs and the Ig-like module are denoted by rectangles and framed if duplicated together. The horizontal lines bordered by slashes may represent segments of identical or distinct chromosomes. See the text for further explanation.

G1L2 shows a higher degree of homology to the L2 module of versican than to G2L2, suggesting gene conversion after separation. In general, L2 modules seem to be less diverse than L3, showing that functional constraints imposed on L2 prevented mutations from being fixed. Modules of the aggrecan G2 domain, which cannot bind LP or hyaluronan [27], do not show significant sequence differences from the others, although the inter-species divergence is slightly higher than that of G1. This finding agrees with the observation by Perkins et al. [3], who have shown that there is no difference in sequence and net charge of the modules of G1 or G2.

On the basis of the sequence similarity and copy number of the HABMs in the various proteins, we suggest a model for their evolution (Figure 2). Following the first duplication, the CD44 module started to evolve separately from the primordial sequence. After another duplication, one copy became the ancestor of the gene for the TSG-6 module, and the other duplicated again to code for the tandem L2 and L3 modules. At this stage the duplicated segment acquired an Ig-like L1 ancestor gene via recombination within phase 1 introns, which had evolved separately up to this point. The genetic element encoding the three modules duplicated further, resulting in segregation of the genes for link protein and proteoglycans, and the latter duplicated twice to form the ancestor genes for neurocan, versican and aggrecan. Finally, genetic elements coding for aggrecan G1 and G2 domains were duplicated and diverged. In summary, altogether seven duplication events were involved in the evolution of the 12 copies of the HABM.

From previous comparisons it was suggested by Neame et al. [8] that the gene duplication which resulted in L2 and L3 arose at a time similar to that of the separation of the genes for LP and aggrecan. Alignment of threefold more sequences convinced us that, although the two duplications were close in time, the common ancestor of L2 and L3 duplicated first.
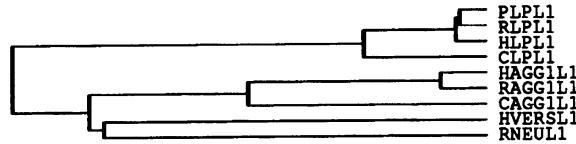

Figure 3   Dendrogram relating the sequences of the Ig-like L1 domains

For further details and abbreviations, see Figure 1.

The model does not specify how many stages the L1 module ancestor has gone through before joining the HABM. It implies, however, that since the time when the L1 ancestor joined the duplicated HABMs, they have evolved together. A dendrogram obtained for the L1 module (Figure 3) confirms that the order of segregation during evolution was LP–neurocan and versican–aggrecan. The same conclusion can be drawn from the dendrogram constructed by Chandrasekaran and Tanzer [12], who compared the L1–L3 modules together.

The rate of divergence for the L1 module is higher than that of L2/L3. The change in the function may be responsible for that. Whereas the L2/L3 modules retained the hyaluronan-binding capacity, the L1 modules of LP and aggrecan have evolved in such a way that they have become capable of binding each other.

The above evolutionary scheme suggests a minimum number of seven for duplications necessary to make the six genes, but additional duplications cannot be excluded. It is possible that duplication of the single ancestor sequence has led to generation of further, as yet undetected, modules before the tandem array was fixed. A priori, it is also possible that the L2–L3 tandem repeat has combined with modules other than L1. Finally, the L1–L3 modules together may have been incorporated as building blocks into other large multi-domain proteins. Discoveries in

future years will presumably supply the missing details to a more complex picture of the evolution of the HABM.

Endre BARTA,*† Ferenc DEÁK*‡ and Ibolya KISS*

*Institute of Biochemistry, Biological Research Center, Hungarian Academy of Sciences, H6701 Szeged, P.O. Box 521, and †Institute for Plant Sciences, Agricultural Biotechnology Center, H2101 Gödöllő, P.O. Box 170, Hungary

‡ To whom correspondence should be addressed.

1   Neame, P. J., Christner, J. E. and Baker, J. R. (1986) J. Biol. Chem. **261**, 3519–3535
2   Deák, F., Kiss, I., Sparks, K. J., Argraves, W. S., Hampikian, G. and Goetinck, P. F. (1986) Proc. Natl. Acad. Sci. U.S.A. **83**, 3766–3770
3   Perkins, S. J., Nealis, A. S., Dudhia, J. and Hardingham, T. E. (1989) J. Mol. Biol. **206**, 737–753
4   Dudhia, J. and Hardingham, T. E. (1990) Nucleic Acids Res. **18**, 1292
5   Bonnet, F., Périn, J.-P., Lorenzo, F., Jollès, J. and Jollès, P. (1986) Biochim. Biophys. Acta **873**, 152–155
6   Périn, J.-P., Bonnet, F., Thurieau, C. and Jollès, P. (1987) J. Biol. Chem. **262**, 13269–13272
7   Goetinck, P. F., Stirpe, N. S., Tsonis, P. A. and Carlone, D. (1987) J. Cell Biol. **105**, 2403–2408
8   Neame, P. J., Christner, J. E. & Baker, J. R. (1987) J. Biol. Chem. **262**, 17768–17778
9   Doege, K. J., Sasaki, M., Horigan, E., Hassell, J. R. and Yamada, Y. (1987) J. Biol. Chem. **262**, 17757–17767
10 Antonsson, P., Heinegård, D. and Oldberg, Å. (1989) J. Biol. Chem. **264**, 16170–16173
11 Doege, K. J., Sasaki, M., Kimura, T. and Yamada, Y. (1991) J. Biol. Chem. **266**, 894–902
12 Chandrasekaran, L. and Tanzer, M. L. (1992) Biochem. J. **288**, 903–910
13 Zimmermann, D. and Ruoslahti, E. (1989) EMBO J. **8**, 2975–2981
14 Rauch, U., Karthikeyan, L., Maurel, P., Margolis, R. U. and Margolis, R. K. (1992) J. Biol. Chem. **267**, 19536–19547
15 Goldstein, L. A., Zhou, D. F. H., Picker, L. J., Minty, C. N., Bargatze, R. F., Ding, J. F. and Butcher, E. C. (1989) Cell **56**, 1063–1072
16 Stamenkovic, I., Amiot, M., Pesando, J. M. and Seed, B. (1989) Cell **56**, 1057–1062
17 Lee, T. H., Wisniewski, H.-G. and Vilcek, J. (1992) J. Cell Biol. **116**, 545–557
18 Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. **12**, 387–395
19 Gribskov, M. and Burgess, R. R. (1986) Nucleic Acids Res. **14**, 6745–6763
20 Nottenburg, C., Rees, G. and St. John, T. (1989) Proc. Natl. Acad. Sci. U.S.A. **86**, 8521–8525
21 Günthert, U., Hofmann, M., Rudy, W., Reber, S., Zöller, M., Haussmann, I., Matzku, S., Wenzel, A., Ponta, H. and Herrlich, P. (1991) Cell **65**, 13–24
22 Idzerda, R. L., Carter, W. G., Nottenburg, C., Wayner, E. A., Gallatin, W. M. and St. John, T. (1989) Proc. Natl. Acad. Sci. U.S.A. **86**, 4659–4663
23 Bosworth, B. T., St. John, T., Gallatin, W. M. and Harp, J. A. (1991) Mol. Immunol. **28**, 1131–1135
24 Aruffo, A., Stamenkovic, I., Melnick, M., Underhill, C. B. and Seed, B. (1990) Cell **61**, 1303–1313
25 Kiss, I., Deák, F., Mestric, S., Delius, H., Soós, J., Dékány, K., Argraves, W. S., Sparks, K. J. and Goetinck, P. F. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 6399–6403
26 Rhodes, C., Doege, K., Sasaki, M. and Yamada, Y. (1988) J. Biol. Chem. **263**, 6063–6067
27 Fosang, A. J. and Hardingham, T. E. (1989) Biochem. J. **261**, 801–809