doi:10.1016/j.jmb.2010.07.011 J. Mol. Biol. (2010) xx, xxx–xxx

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Nebulin: A Study of Protein Repeat Evolution

Åsa K. Björklund, Sara Light†, Rauan Sagit† and Arne Elofsson*

Center for Biological Membrane Research, Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

Received 30 March 2010; received in revised form 22 June 2010; accepted 7 July 2010

Edited by M. Levitt

Protein domain repeats are common in proteins that are central to the organization of a cell, in particular in eukaryotes. They are known to evolve through internal tandem duplications. However, the understanding of the underlying mechanisms is incomplete. To shed light on repeat expansion mechanisms, we have studied the evolution of the muscle protein Nebulin, a protein that contains a large number of actin-binding nebulin domains.

Nebulin proteins have evolved from an invertebrate precursor containing two nebulin domains. Repeat regions have expanded through duplications of single domains, as well as duplications of a super repeat (SR) consisting of seven nebulins. We show that the SR has evolved independently into large regions in at least three instances: twice in the invertebrate *Branchiostoma floridae* and once in vertebrates.

In-depth analysis reveals several recent tandem duplications in the Nebulin gene. The events involve both single-domain and multidomain SR units or several SR units. There are single events, but frequently the same unit is duplicated multiple times. For instance, an ancestor of human and chimpanzee underwent two tandem duplications. The duplication junction coincides with an Alu transposon, thus suggesting duplication through Alu-mediated homologous recombination.

Duplications in the SR region consistently involve multiples of seven domains. However, the exact unit that is duplicated varies both between species and within species. Thus, multiple tandem duplications of the same motif did not create the large Nebulin protein.

Finally, analysis of segmental duplications in the human genome reveals that duplications are more common in genes containing domain repeats than in those coding for nonrepeated proteins. In fact, segmental duplications are found three to six times more often in long repeated genes than expected by chance.

Keywords: protein domain repeat; evolution; repeat duplication; segmental duplication; Nebulin

Introduction

The creation of new multidomain architectures through shuffling of protein domains is an important mechanism that serves to expand the protein repertoire and has been studied extensively during the last few years.[1–4] However, one type of domain rearrangement has often been ignored: the creation of protein segments consisting of several domains of the same type in tandem–domain repeats.

Domain repeats are present in all kingdoms of life and are particularly common in multicellular organisms.[1,5,6] These repeats have been proposed to provide eukaryotes with an extra source of variability to compensate for low generation rates.[7] In addition, long repeats have been shown to be exceptionally common in proteins that are involved in many protein–protein interactions.[6,8]

A probable mechanism for the development of domain repeats is via tandem duplications within a

*Corresponding author. E-mail address: arne@bioinfo.se.

† S.L. and R.S. contributed equally to this work.

Abbreviations used: SR, super repeat; NEU, nebulin evolutionary units.

0022-2836/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.

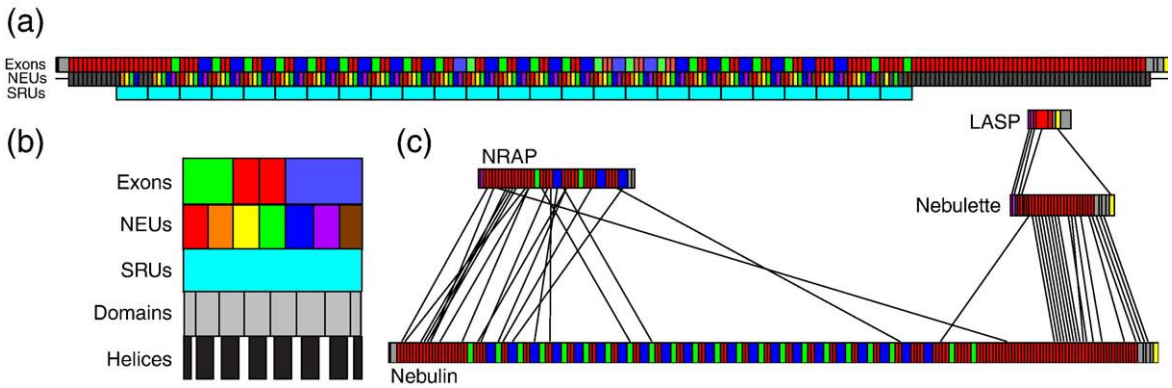

Fig. 1. Overview of human Nebulin genes. (a) Gene structure for human Nebulin. The top line represents nebulin-containing exons with 1 NEU exon (red), 2 NEU exons (green), and 3 NEU exons (blue), or SH3 exons (yellow), LIM exons (magenta), and other exons (gray). The second line represents the 238 evolutionary units in Nebulin, with the SR units shown (in order) in red, orange, yellow, green, blue, magenta, and black, while gray NEU cannot be classified into either of the SR types. The bottom line shows turquoise boxes illustrating each SR unit (SRU). (b) Detailed view of an SR illustrating in what order the SR NEU have been classified, with domain boundaries (as defined by Pfam) in light gray; at the bottom, the secondary structure prediction is shown with black boxes for regions with predicted α helices. (c) Overview of the four human Nebulin proteins (Nebulin, Nebulette, NRAP, and LASP) showing lines for reciprocal best hits alignments between the individual exons; exon coloring as in (a).

gene,[9] where a segment is duplicated and the copy is inserted next to its origin. However, the exact mechanism behind this phenomenon is not fully understood. In a previous study,[6] we examined in detail internal duplication in domain repeats of higher eukaryotes and found that it is similar to duplication of multiple domains all at once. The size of the duplication units varied both within domain families and between domain families.

In the case of the muscle protein Nebulin, we found that a region containing seven nebulin domains seems to have been duplicated repeatedly in a clear pattern.[6] The seven-domain motif of Nebulin is what we refer to as a super repeat (SR). Upon further investigation of the Nebulin gene, we found clear patterns of both SR duplications and duplications of single nebulin domains. Due to its clear, yet diverse, duplication pattern, this protein was selected for further investigation of the mechanisms underlying the evolution of protein repeats.

The nebulin domain is an approximately 35-residue-long helical structure that is involved in actin binding.[10] It contains a conserved SDxxYK sequence pattern that is responsible for binding to actin.[11] In some contexts, nebulin domains form an SR consisting of seven domains. The SR motif has a higher affinity for actin than single nebulins.[12] In addition, the SR contains a conserved WLKGIGW motif in every seventh unit that is thought to be responsible for binding to the thin filament components tropomyosin and troponin.[11] The functional advantage of having this seven-domain module could explain its common duplication.

In vertebrates, there are four protein families that contain the nebulin domain (Fig. 1c); the most well studied is the giant protein Nebulin. The protein is mainly found in skeletal muscle where it is implicated in thin filament formation. In Nebulin, there is a clear SR pattern, with more than 20 SRs flanked by simple repeat nebulin domains and an SH3 domain at the C-terminus.[13] The Nebulin protein undergoes alternative splicing, which regulates its length, thereby affecting the size of muscle thin filaments.[11]

NRAP (*N*ebulin-*r*elated *a*nchoring *p*rotein) is the only other protein that contains nebulin SRs, although only five copies. NRAP is found in both skeletal and cardiac muscles and is believed to be involved in myofibril assembly.[11] LASP (*LIM a*nd *S*H3 *p*rotein) is a short nebulin-containing protein that is present in both invertebrates and vertebrates. LASP is expressed in most tissues and seems to be involved in cytoskeletal organization.[11] Finally, the Nebulette protein consists of a nebulin repeat, but does not have SRs like NRAP and Nebulin. Although it is predominantly found in cardiac muscle, where it is probably involved in myofibril assembly,[14] Nebulette also has a ubiquitously expressed splice form with the same domain structure as LASP.[15,16]

Recently, the first invertebrate Nebulin protein was discovered in the amphioxus cephalocordate *Branchiostoma floridae*.[17] The gene structure of this newly discovered invertebrate protein, whose relatives generally are found in skeletal muscle, is vastly different from the structures of vertebrate genes. In fact, it consists of two regions of SRs and, additionally, a third region of simple repeats covering a single large exon.

Here, we first present an analysis of the evolution of the nebulin-containing genes and the creation of

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

the nebulin SR. Subsequently, we present an in-depth study of recent tandem duplications in several vertebrate Nebulin orthologs. Finally, we assess the prevalence of protein domain repeats in segmental duplications of the human genome.

Results

Domains and exons in nebulin genes

At the start of this investigation, Pfam domains[18] were assigned to the nebulin-containing genes. The nebulin domain consists of an α helix in its actin-bound form, and the nebulin molecule is believed to extend along the myofibril since the length of one nebulin unit in helical conformation corresponds to the size of one actin monomer.[19] However, the conformation of nebulin in its unbound form is not known. Nebulin units are transient helices in solution,[10] while long constructs of nebulin repeats aggregate at natural pH.[20] Attempts to model the structure of nebulin domains were not successful, since it lacks sufficiently close homologs in structure databases. Furthermore, *de novo* protein structure prediction on a nebulin fragment, using Rosetta,[21] did not provide any clear structural models.

When exon boundaries within repeats and SRs were examined, it became evident that they do not coincide with the domain boundaries as defined by Pfam. The exon boundaries were mainly located in the middle of a nebulin domain in Nebulin (Fig. 1b), as well as in other nebulin-containing proteins. The nebulin domain is the functional unit that most likely corresponds to an α helix, according to predictions (Fig. 1b), However, during extension by tandem duplications, an exon consisting of two partial helices—not the domain as defined by Pfam—is the evolutionary unit. Hence, we have chosen to disregard the Pfam domain definition and instead consider "Nebulin Evolutionary Units" (NEU) defined by the exon boundaries (Fig. 1b).

Thus, we used exons as the basis for our assignment of NEU in proteins from other species. To avoid annotation bias due to insufficient mRNA sampling, we ran annotations on the full DNA sequences of all nebulin genes using profiles created from exons in human and mouse nebulin genes. With this method, we found several new exons that were unannotated in Ensembl. Naturally, we cannot say anything about the expression of such exons; however, we are mainly interested in evolutionary events that have occurred within the gene. Thus, the presence of an NEU at a certain position is sufficient for our purposes. In the human Nebulin gene, we identified 13 new exons that were not included in Ensembl. We are still confident that most are coding exons, since a region composed of 11 of these exons was identified by Donner *et al.* in a minor splice form.[13]

Evolution of the nebulin gene families

Based on our assignments, the human Nebulin protein contains 238 NEU, followed by an SH3 domain. A summary of the nebulin genes in all species can be found in Table 1. Clearly, all vertebrates contain long Nebulin proteins. However, in some cases, the Nebulin gene has been truncated or has not been completely sequenced. Such fragmented proteins have not been included here. Several of the less well-sequenced vertebrates have no identified LASP, while most of the fishes have two LASPs, indicating gene duplication in the fish lineage. The only invertebrates in Ensembl that have long nebulin repeats are *Ciona savignyi* and *Ciona intestinalis*, with 36 and 41 NEU, respectively.

LASP contains a LIM domain at the N-terminus and an SH3 domain at the C-terminus, as well as two or three nebulin domains. It is found in at least one copy in almost all vertebrates and invertebrates, even in insects (Table 1). Therefore, it is likely that the evolution of the nebulin-containing proteins started with a LASP-like protein.

Subsequently, the LASP-like precursor to Nebulette, NRAP, and Nebulin must have evolved through several tandem duplications of single nebulin domains. At some point in evolution, the first SR was then formed in a precursor to Nebulin and NRAP. Since the SR motif has a higher affinity for actin filaments,[12] it was likely advantageous to copy the whole motif rather than single nebulin domains.

Nebulin and Nebulette still have the C-terminal SH3 domain; in addition, Nebulette has an exon coding for an N-terminal LIM domain that is only included in the LASP-like isoform of the Nebulette gene.[16] NRAP, on the other hand, has an N-terminal LIM domain, but no SH3 domain. Examining the sequence similarity, we found that there is a clear resemblance between the C-terminal regions of Nebulin and Nebulette, and between the N-terminal parts of Nebulin and NRAP (Fig. 1c). This finding further supports a common origin for the three vertebrate nebulin-containing proteins.

The most probable evolutionary scenario involves a precursor with both SRs and simple repeats, and the LIM and SH3 domains. This precursor must have lost different regions in three different paralogs and undergone duplications of the SR regions in NRAP and Nebulin. This would explain the similarity of Nebulin to C-terminal Nebulette and N-terminal NRAP.

Creation of the SR

In the human Nebulin gene, each SR containing 7 NEU occurs in 25 copies, based on our assignments. Even though each NEU in the SR differs markedly from the other six and from other non-SR NEU,

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

Table 1. Overview of nebulin-containing genes in selected species, with protein length (P_{len}) and the number of nebulin exons for all four Nebulin protein families

		Nebulin		NRAP		Nebulette		LASP	
Species	Common name	P_{len}	Exons	P_{len}	Exons	P_{len}	Exons	P_{len}	Exons
Homo sapiens	Human	8556	181	1737	42	1167	35	478	8
Pan troglodytes	Chimpanzee	8455	180	1737	42	1113	32	260	7
Gorilla gorilla	Gorilla	7315	173	1665	44	1006	30	260	7
Pongo pygmaeus	Orangutan	2914	78	1728	43	1066	30	310	8
Microcebus murinus	Mouse lemur	7232	166	1734	43	962	30	340	13
Otolemur garnettii	Bushbaby	6624	172	1728	46	1007	33	285	12
Ochotona princes	Pika	7540	200	1725	47	940	38	260	8
Mus musculus	Mouse	7526	163	1728	42	1154	33	268	7
Rattus norvegicus	Rat	7423	165	1727	42	1143	35	272	10
Cavia porcellus	Guinea pig	7519	166	1732	45	369	16	239	6
Bos taurus	Cow	7735	170	1746	44	1011	29	259	7
Tursiops truncatus	Dolphin	7833	176	1723	46	1006	32	254	12
Vicugna pacos	Alpaca	7704	200	1727	50	1033	36	271	5
Canis familiaris	Dog	791	26	1735	42	1012	28	395	13
Felis catus	Cat	7163	179	1733	44	1005	31	237	6
Equus caballus	Horse	8075	172	1724	42	1010	28	239	6
Myotis lucifugus	Microbat	7365	169	1509	38	938	26	349	14
Pteropus vampyrus	Megabat	7622	185	1738	43	997	33	259	7
Erinaceus europaeus	Hedgehog	7098	185	1793	50	973	28	352	17
Procavia capensis	Hyrax	7500	176	1687	46	1004	33	263	7
Loxodonta africana	Elephant	7913	210	1721	48	1003	29	260	10
Echinops telfairi	Lesser hedgehog tenrec	7353	191	1729	45	939	28	171	10
Choloepus hoffmanni	Sloth	7442	196	1717	54	947	32	150	4
Dasypus novemcinctus	Armadillo	7451	203	1668	43	1001	35	238	7
Macropus eugenii	Wallaby	7151	180	1725	43	1024	31	239	8
Monodelphis domestica	Opossum	7577	166	1737	42	1156	33	260	8
Ornithorhynchus anatinus	Platypus	8342	191	1731	42	1174	34	86	3
Gallus gallus	Chicken	7043	155	1595	38	1063	33	242	8
Anolis carolinensis	Anole lizard	825	24	1679	42	1036	27	262	8
Xenopus tropicalis	*X. tropicalis*	6066	135	—	—	—	—	307	13
Tetraodon nigroviridis	Tetraodon	—	—	1734	43	1462	34	231	10
								238	7
Takifugu rubripes	Fugu	4545	102	1769	45	254	7	283	8
Oryzias latipes	Medaka	6513	146	1765	45	1118	30	266	7
								272	8
Gasterosteus aculeatus	Stickleback	6039	124	1739	44	1015	32	256	8
								233	7
Danio rerio	Zebrafish	5444	116	—	—	—	—	—	—
Branchiostoma B. floridae	Florida lancet	16,125	274	—	—	—	—	—	—
C. savignyi	*C. savignyi*	—	—	1629[a]	36	—	—	282	12
								237	10
C. intestinalis	*C. intestinalis*	—	—	1707[a]	41	—	—	337	13
								311	12
Pyretophorus gambiae	Anopheles	—	—	—	—	—	—	182	4
Drosophila melanogaster	Fruitfly	—	—	—	—	—	—	193	5

Species are arranged according to the Ensembl species tree, and some species were removed due to clearly fragmented Nebulin genes. Still, several of the genes in this list have incomplete sequence coverage or fragmented gene predictions; hence, the size of the Nebulin genes and the number of exons may be underestimated.

[a] The two Ciona proteins are classified as orthologs to NRAP according to Ensembl; however, they contain an SH3 domain and have higher alignment scores to Nebulin.

there is fairly good sequence conservation within each type of SR NEU. For instance, the typical WLKGIG motif can clearly be seen in SR unit 1 and is unique compared to the other 6 NEU (see Supplementary Material).

The number of NEU per exon within the SR varies. Most frequently, they are contained within 4 exons: one of intermediate size is equal to 2 NEU, two short ones correspond to 1 NEU, and a large one contains 3 NEU (Figs. 1a and 2b). The N-terminal and C-terminal SR units consist of mainly single exons, and some repeats have lost one or two introns.

Interestingly, in NRAP, we only see 2 NEU and 3 NEU exons in the SRs, and no instance of the 2,1,1,3 SR structure found in Nebulin (Fig. 2b). These findings indicate that the SR of seven domains was, from the beginning, contained in seven separate exons. It is possible that two introns had been lost before the gene duplication that separated NRAP and Nebulin. In Nebulin, two other introns must

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

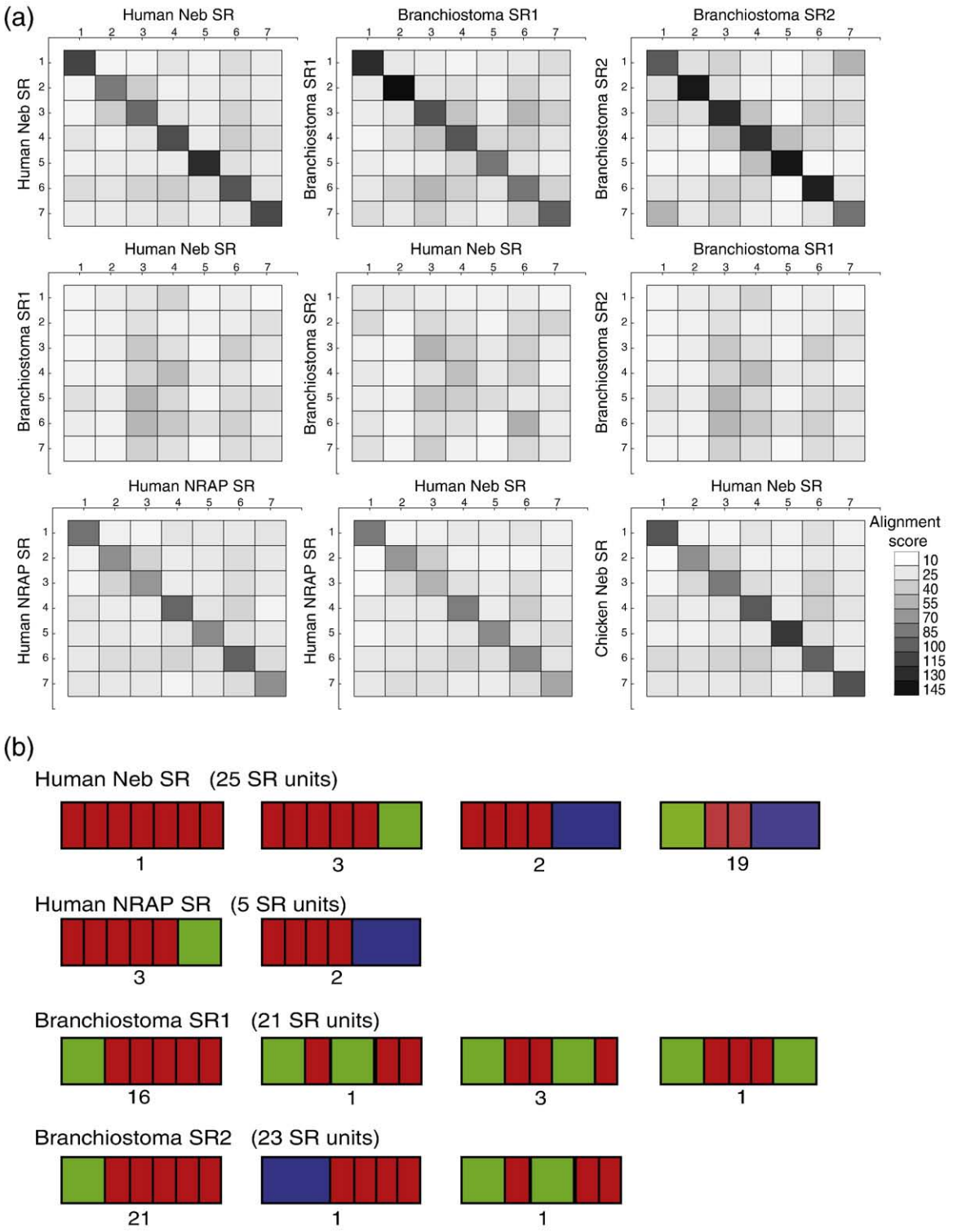

Fig. 2. (a) Similarity matrices based on the average alignment score among all the NEU in the human Nebulin (Neb), human NRAP, chicken Nebulin, and two *B. floridae* SRs. (b) Intron losses in the SRs. The distribution of different exon patterns in SRs is illustrated as boxes, with the same color scheme as in Fig. 1. The number below each SR is the number of SRs with the specific exon structure.

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

have been lost before the expansion of most of the SR structure.

In addition to vertebrate Nebulin and NRAP, the Nebulin of the Florida lancelet *B. floridae* contains two distinct regions with SRs. The gene and the NEU structure of the lancelet Nebulin can be found in Supplementary Material. Upon a more detailed examination of these different SR regions, it was evident that the vertebrate Nebulin and NRAP SRs are homologous, as can be seen in Fig. 2a. However, the two lancelet SRs are not similar to each other or to the two vertebrate SRs (Fig. 2a). Also, the NEU of the human Nebulin SR are more similar to their chicken counterpart than to the NEU of the human NRAP SR.

Furthermore, clustering of Nebulin NEU from SRs in human and lancelet indicates a low level of similarity between the different units of each SR, as they do not cluster together (Supplementary Material). In addition, there is a clear difference in the exon/intron patterns of vertebrate and *B. floridae* SR (Fig. 2b). Naturally, it is hard to define the exact order of the *B. floridae* SR units. Here, they were ordered in a way that would maximize similarity to human Nebulin SR along the diagonal in Fig. 2a. Hence, we cannot exclude the possibility that the loss of an intron between the first NEU and the second NEU in the lancelet SR is the same as that between the sixth NEU and the seventh NEU in the human SR.

Furthermore, it is worth noting that the lancelet SR lacks the typical WLKGIGW motif. Although the lancelet SR may bind troponin/tropomyosin through other means, lack of the motif casts some doubt on the assumption that the lancelet SR has the same function as its vertebrate orthologs.

In consideration of all these observations, it appears that the seven-domain motif has been duplicated in large regions at least three independent times in evolution: once in a vertebrate precursor and twice in the lancelet. Indeed, we cannot exclude the possibility that the 7-NEU SR motif was created two or three independent times.

The Nebulin gene in vertebrates

Even though most vertebrate Nebulin genes are highly similar, the number of NEU varies between species, and we have found several examples of recent duplications in different lineages. However, many of the duplication events are so old that it is difficult to elucidate the exact duplication break points. Still, the number of SR units in each organism and their multiple alignments give an indication about the location and the nature of the duplication events. We estimate that most duplications occurred in an early vertebrate, as we see a common pattern of 16 SRs in all sequenced vertebrates. Through manual inspection, we have identified a number of independent events that are illustrated along the branches of a species tree for a selection of well-sequenced organisms in Fig. 3. The largest Nebulin proteins are found in human, chimpanzee, and platypus with 25 SRs, while the mouse protein has the same SR structure as the mammalian ancestor, and nonmammalian species have fewer SR units than mammals.

The SR duplications in the mammalian branch of the tree were studied in particular detail (summarized in Table 2). Figure 4 shows an overview of the multiple alignment of the Nebulin gene products for some mammals where several events are evident. For instance, there is a duplication of several exons in human and chimpanzee in the middle of the protein. Other events in the SR region include two duplications in platypus, one duplication in cow, one duplication in elephant, and one duplication in horse. In addition, we found several duplications in the C-terminal region of Nebulin involving 1 exon—or in some cases several exons—in the different species.

It should also be mentioned that gaps in the alignment for single species may reflect missed assignments. One such example is found in the duplicated region of chimpanzee and human where

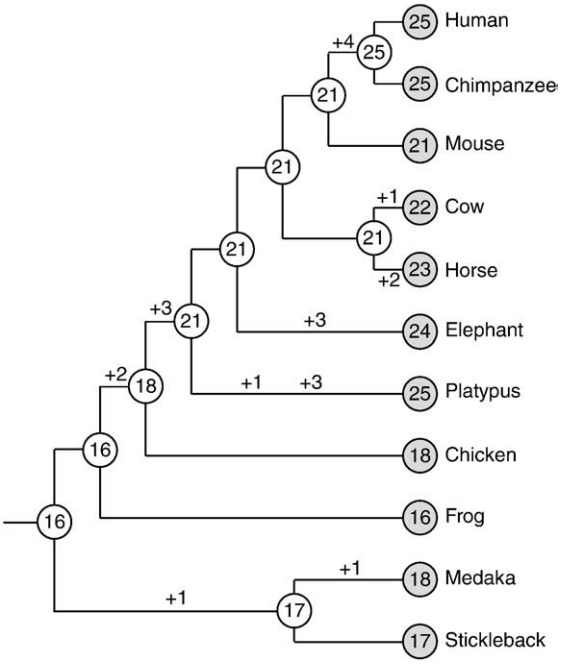

Fig. 3. SR duplications in vertebrates. We estimate that a common ancestor to all the sequenced vertebrates contained 16 SR units. Some of the documented duplication events along different branches are illustrated in the tree for a selection of well-sequenced organisms. The numbers at each node/leaf correspond to the assumed numbers of SRs, and each duplication event is illustrated along the different branches. The branch lengths of the species tree do not reflect any evolutionary distances.

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

Table 2. Summary of documented duplication events in vertebrate Nebulin (numbers in the first column refer to the events illustrated in Fig. 4)

Number	Species	Size (kb)	Times	NEU	SR order	DNA repeats
1	Human/chimpanzee	10.5	2	14	1, 2, 3, 4, 5, 6, 7	Alu (SINE)
2	Human	0.7	2	1–2	C-terminal[a]	No
3	Chimpanzee	0.7	1	1	C-terminal[a]	No
4	Cow	3.8	1	6 (7)	4, 5, 6, 7, 1, 2 (3)	LINE-1
5	Horse	10.3	1	14	1, 2, 3, 4, 5, 6, 7	No
6	Horse	1.1	3	1	C-terminal[a]	No
7	Elephant	5.5	3	7	5, 6, 7, 1, 2, 3, 4	AfroSINE
8	Elephant	0.7	3	1	C-terminal[a]	No
9	Platypus[b]	10+5	1.5	14+7	6, 7, 1, 2, 3, 4, 5	DNA-hAT Blackjack[c]
10	Platypus	7.4	1	7	3, 4, 5, 6, 7, 1, 2	LINE-2[c]
11	Platypus	1.4	1	1	C-terminal[a]	LINE-2[c]
—	Chicken	0.9	4	1	C-terminal[a]	No

For each event, the genomic size duplicated (in kb), the number of NEU, the number of times the unit is duplicated, and the SR order duplicated are presented. The last column indicates whether there are any interspersed DNA repeats identified at the duplication boundaries.

[a] A duplication in the simple repeat region close to the C-terminus.

[b] A duplication of one large region and one smaller region (covering half of the larger one).

[c] The DNA repeat is found within the duplicated region; hence, the occurrence at multiple break points may be a consequence of tandem duplication.

there is a gap for 1 exon in chimpanzee. The intron between the surrounding exons has a large unsequenced region; hence, that the exon in question has been missed in the assignments, rather than genuinely deleted, is more likely, since the latter would disrupt the SR structure.

Double-tandem duplication in the human/chimpanzee SR

Alignment of the exons and the DNA sequence of human Nebulin shows a region of high internal symmetry (Fig. 5, squares) where a region has been duplicated twice. The exact same pattern can be seen in the chimpanzee protein, but not in any other species, not even in other primates. This region corresponds to two documented segmental duplications according to the Human Segmental Duplication Database.[22] The duplicated region spans about 10.5 kb and contains 8 exons starting with a 2-NEU exon. Hence, in this case, the unit of SR evolution is two motifs, with the exons containing 2, 1, 1, and 3 NEU. In this article, we define an SR as we identified it in this event, where the first NEU in the SR corresponds to the first NEU in that exon (Fig. 1b). However, the exact unit being duplicated may not always be identical to this case, as we will show later on.

The three human regions have over 99% sequence identity to one another, as well as to their chimpanzee counterparts, and it is impossible to distinguish any pair as more similar, leading to the conclusion that the two duplications occurred at approximately the same time point. Further inspection indicates that the break-point region partly overlaps with a predicted Alu repeat.[23] It is possible that the homology between Alu elements was responsible for the homologous recombination leading to the tandem duplication of that region. In addition, there are remnants of a LINE-2 element located in the break-point exons surrounding the duplicated regions. Donner *et al.* suggested that this LINE-2 could be responsible for the duplication event.[13]

It should be noted that the duplicated region is not part of a major splice form and that it has only been seen in fetal transcripts.[13] Thus, the functional implications of this duplication are unclear. In addition, the region that has been duplicated is associated with copy number variations based on three different experiments. In one case, a loss of part of the region has been reported,[24] while in two other studies, the region is associated with copy number gain.[25,26] However, there is not enough information to determine if the boundaries of these copy number variations coincide exactly with the boundaries of the tandem duplication documented here. Still, this suggests that tandem duplication of these human SR units is an ongoing evolutionary process, and this may be the case in other species as well.

Other mammalian SR duplications

Upon inspection, we found several other examples of recent tandem duplications in the SR region of Nebulin. These events can be seen in the multiple alignments in Fig. 4, where each unit of duplication has been marked with a box, and events are summarized in Table 2. There, we have only taken into account very recent duplications with at least a 90% sequence identity in both noncoding and coding regions.

One interesting example is the platypus Nebulin, which has tandem duplications in two parts of the

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

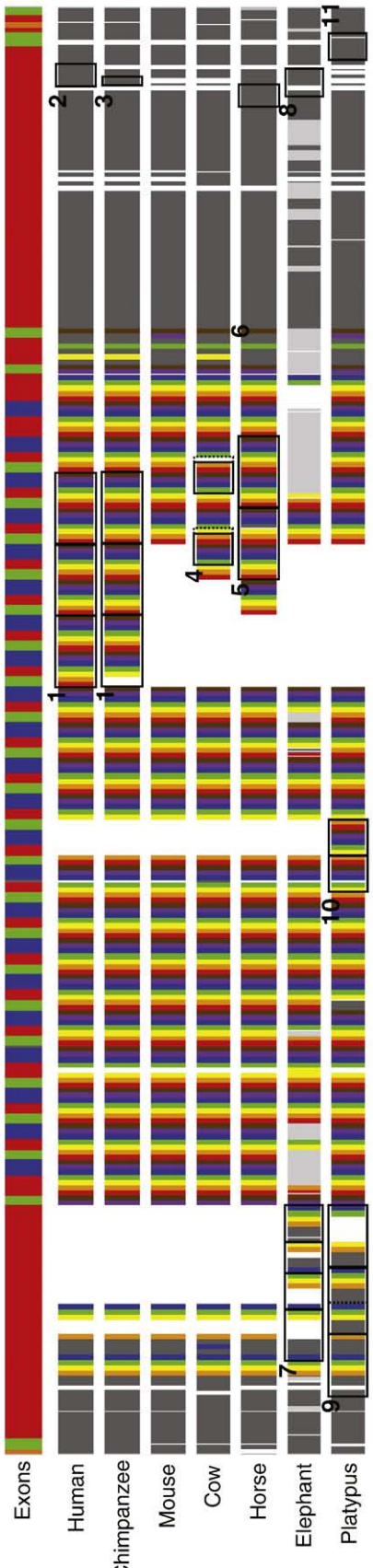

Fig. 4. Overview of multiple alignment with NEU coloring (with the same color scheme as in Fig. 1). Unsequenced regions are shown in light gray, and white spaces represent gaps in the sequence alignment. Above the alignment is the exon structure with coloring according to 1 NEU exon (red), 2 NEU exons (green), and 3 NEU exons (blue). A multiple alignment of the sequences can be found in Supplementary Material. Each documented event that is mentioned in the text has been highlighted with a box around each homologous region. The duplication of cow also includes a seventh NEU (broken lines) that corresponds to a sequencing gap in one of the copies. In the case of the C-terminal events, all exons affected by the duplications are within one box. The numbers by each box are used to refer to each event in Table 1.

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

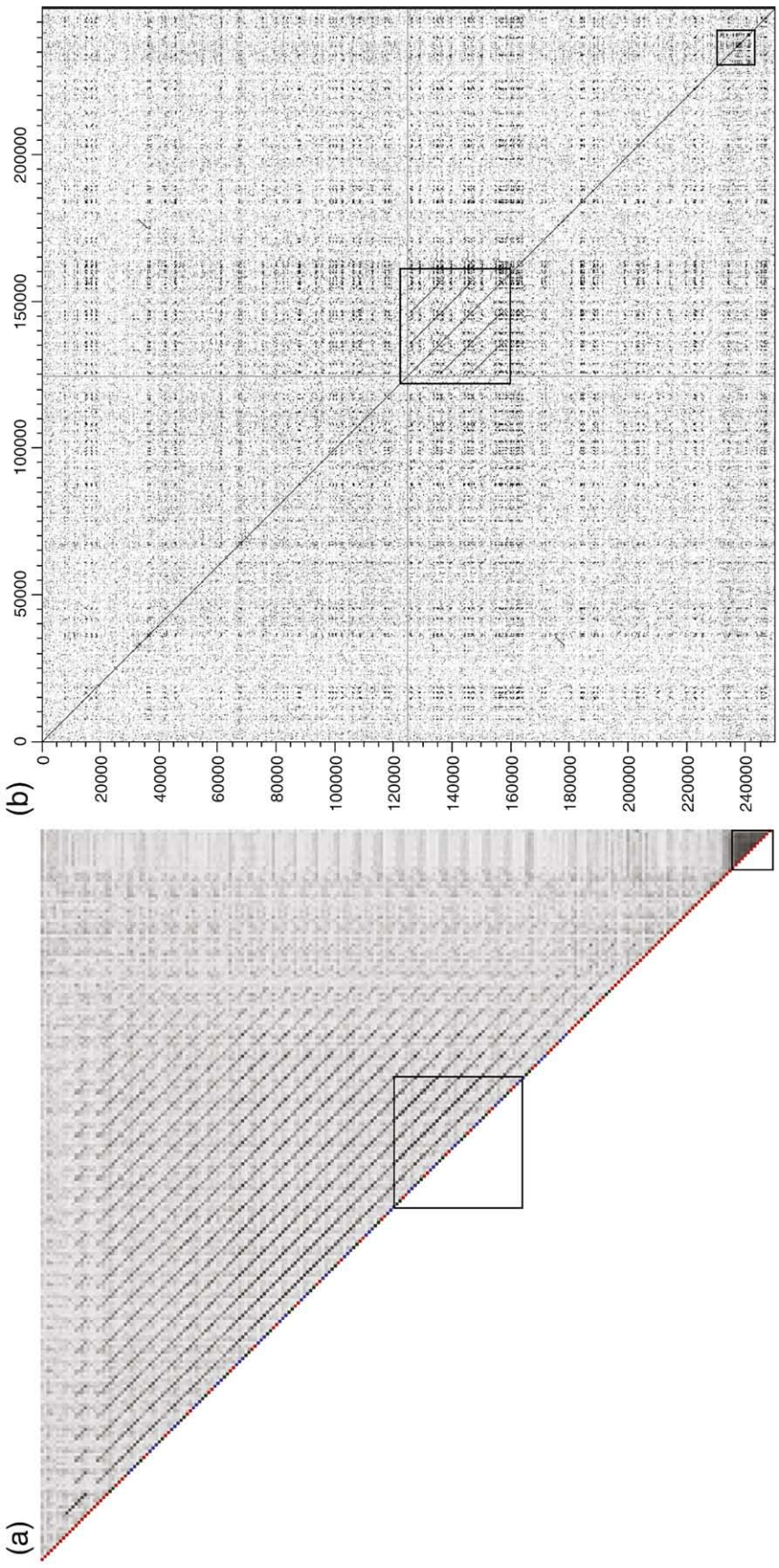

Fig. 5. (a) Similarity matrix for the human Nebulin protein, where each column/row represents an NEU and color intensity reflects the alignment score between two NEU; the diagonal is colored according to exon pattern, in the same color scheme as Fig. 1. (b) Dotplot for the human Nebulin gene. DNA alignment with sliding window, where regions with a high level of similarity are highlighted with dark fields. Boxes in both plots illustrate examples of recent tandem duplication: one in the SR region and one close to the C-terminus.

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

SR. There are two large duplications close to the N-terminus (with one spanning about 10 kb) that seem to contain 14 exons. Another duplication involving half of the larger duplication unit (i.e., 7 exons) has also occurred. However, it is unclear which of the two duplications happened first; it is likely that they were duplicated roughly at the same time in evolution. The other duplication in platypus is found in the middle of the SR and is a duplication of 4 exons containing 7 NEU. Even though the number of NEU is a multiple of 7, these duplications do not contain the same SR unit as the human/chimpanzee event. In the first example, the region that is duplicated starts with the sixth NEU in the SR; in the second case, it starts with the third NEU.

The elephant gene is not fully sequenced throughout the gene; however, the N-terminal region has fairly good coverage. In a similar region as the first platypus duplication, a region containing 7 exons appears to have been duplicated three times in the recent past. Two of the four nearly identical regions contain 7 exons with an SR of seven domains. However, there are sequencing gaps in the two middle duplicates; hence, the true number of exons in each unit is uncertain.

In horse Nebulin, there has been tandem duplication in a region close to the event in human/chimpanzee. However, even if this duplication also involves a region with the defined SR order, it is not the same as in the human example. Here, the unit that was duplicated was shifted one SR towards the C-terminus (Fig. 4, event 5). In cow, there is yet another duplication close to the C-terminus of the SR region. The cow duplication is not a true tandem duplication, but rather a duplication into a nonadjacent neighboring region. The two regions contain only 3 exons containing 6 SR NEU. However, one of the regions is ended by a sequence assembly gap (Fig. 4, broken lines); hence, it is a likely duplication of a whole SR.

In conclusion, we see a variety of SR duplications involving different regions and different units of the SRs. In fact, duplication break points are found in as many as five different introns of the SR (Table 2). This indicates that tandem duplications can occur almost at random within a large repeat, as long as the repeat structure is not interrupted.

Expansion of the C-terminal region of Nebulin

The C-terminal region of Nebulin is composed of simple repeats of single NEU exons (Fig. 5a). There are many duplication events in this region. For instance, in the lower right region of Fig. 5, there is similarity between closely situated exons in the C-terminal region of the human Nebulin gene. Inspection of this region at the DNA level reveals three internally highly similar segments, each containing about 1 exon. Similar duplication events have occurred in several organisms at approximately 12–18 exons from the C-terminus of their respective proteins, some of which are marked in Fig. 4.

Although recent duplications have occurred in the C-terminal region of Nebulin in most vertebrates, they are not found in all organisms, such as mouse for instance. Hence, the duplications either occurred in a common ancestor and were subsequently lost in some organisms, or took place independently of one another. The latter hypothesis is supported by the finding of a >90% identity in the noncoding parts of the homologous regions within each species. Such high conservation would not be expected over large evolutionary times (see Methods). Furthermore, multiple-sequence alignments of the C-terminal regions reveal that the events are not entirely overlapping (Fig. 4; Supplementary Material). Thus, we can conclude that the events are likely to be phylogenetically unrelated.

The recently duplicated C-terminal region in the human Nebulin gene, located inside the Z-disc,[27] is involved in binding desmin and is also a region of Nebulin that is known to be involved in different isoforms expressed in different muscle tissue types.[28] Although the location of the alternative splicing sites of nebulin is not known for most of the organisms, our results indicate that some organisms have undergone acquisition of new alternatively spliced exons in the recent past.

Tandem duplications in the human genome

Next, all documented events in the Human Segmental Duplication Database[22] were examined to see how common similar tandem duplications occur in genes other than Nebulin. We found 646 examples of tandem duplications after allowing for a 1-kb gap between the two copies. We found 41 examples of duplications involving protein coding exons, even though they are rare.

Furthermore, we investigated if duplications are more common in genes with domain repeats, as defined by Pfam-A domains.[18] In fact, we found that more than half of the tandem duplications affecting coding regions involve genes that code for proteins with domain repeats. Still, only 21% of all proteins contain repeats, and their genes only cover 36% of the genic regions in the genome. Hence, there is a clear overrepresentation of tandem duplications in genes coding for protein domain repeats.

To evaluate this further, we divided the genes into groups based on the number of domains in the repeats. For those groups, the number of tandem duplications and the expected number of events based on the size of the genes were calculated. Using these values, we evaluated the overrepresentation of tandem duplications in protein domain repeats (Fig. 6). It is clear that proteins without domain repeats have fewer tandem duplications than

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

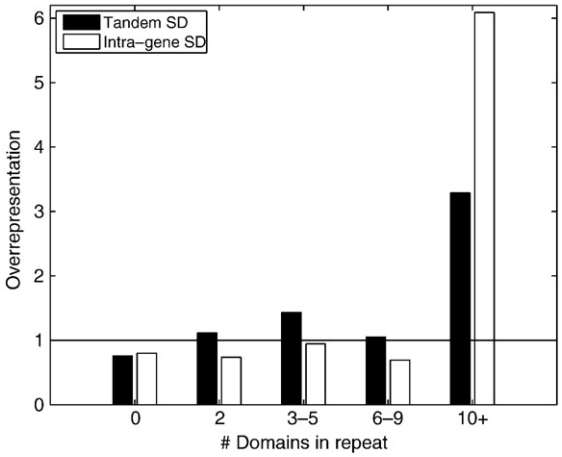

Fig. 6. Segmental duplications (SDs) in domain repeats. For each group of genes with no domain repeats or a different number of domains in the repeats, the bars represent the fraction of SDs that are found, divided by the expected fractions based on their genome coverage. Bars above the line at 1 indicate overrepresentation, while values below indicate underrepresentation. The black bars show tandem duplications with less than 1 kb between the two copies, while the white bars show all SDs that are found within a gene.

expected, while they are overrepresented in all groups of repeated genes. In fact, genes that contain domain repeats of length 10 domains or more have greater than three times more events than expected.

In addition, when all intragene duplications are considered (not only those that are directly in tandem), it is very clear that the proteins with long ($\geq$10 domains) repeats are highly overrepresented. In fact, 14 of the 62 examples of intragene duplications are found in genes with long repeats.

Upon manual inspection of the examples, we found that many of the recent tandem duplications in human domain repeats involved multiple duplications of large cassettes containing several domains (data not shown). Hence, similar types of duplications are seen in those genes as compared to the Nebulin gene.

Summary

The Nebulin gene has evolved through several duplications of nebulin repeat units—either as single nebulin domains or as part of SRs, where a motif of seven nebulins is duplicated together. Our findings indicate that the Nebulin gene has evolved from a protein similar to LASP that was present in the common ancestor of vertebrates and invertebrates. The two nebulin domains in LASP were duplicated several times, and an SR of seven domains was formed. Upon several further duplications of the SR and simple repeat domains, together with loss of the SH3 or LIM domain, the nebulin-containing proteins Nebulin, NRAP, and Nebulette evolved.

In addition, we found that the SR has been expanded independently into large regions at least three times in evolution: once in vertebrates and twice in the invertebrate *B. floridae*. It is clear that several duplications occurred in early vertebrate or prevertebrate evolution, since most of the nebulin repeat units are shared among all vertebrates. However, there is very little evolutionary trace left from such old events.

Nevertheless, we have found several examples of recent duplications of different units of nebulin domains, both in the SR regions and in the simple repeat regions. In some cases, we find single events, but in many instances, we see that the same region has been duplicated multiple times at approximately the same time point. In addition, all duplications in the SR region entail domains in multiples of 7. Still, it is clear that the unit that has been duplicated is not the same in different species.

Furthermore, the duplications in the Nebulin gene are not at all unique, but rather represent a common feature in genes that code for protein domain repeats. In fact, we show that tandem duplications are clearly more common in repeat genes than in genes that do not code for domain repeats.

Discussion

The mechanism behind these tandem duplications is not easy to determine. During the last few years, it has become evident that segmental duplications have had a tremendous impact on the structure of our genomes.[29] Several of these segmental duplications are found in tandem, and many cover gene regions.

Any region of homology between two sequence regions may cause homologous recombination and subsequent tandem duplication.[29] Different types of repeats, such as LINEs (*l*ong *i*nterspersed *n*uclear *e*lement) and SINEs (*s*hort *i*nterspersed *n*uclear *e*lement), may serve as duplication "hot spots" due to their abundance in the genome.[29] As a matter of fact, 27% of human segmental duplications are flanked by the most common human DNA repeat, the Alu repeat.[30] Indeed, we observe that one of the duplicated regions in human Nebulin is flanked by an Alu repeat. However, most of the duplication events in other organisms do not coincide with detected DNA repeats.

It has been shown that tandem segmental duplications were more common in early vertebrate evolution, while interspersed segmental duplications have been dominating the primate lineages.[29] In the case of Nebulin, we see clearly that most

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

duplications occurred in early vertebrates, and that only a few events have taken place in the current species. We have previously shown that the total amount of domain repeats is fairly constant in all vertebrates,[6] which could be explained by this change in segmental duplication mechanism.

Marcotte *et al.* found that the longer is a protein repeat, the more likely it is to be expanded further.[7] It is easy to imagine that once a duplication has occurred, it is also likely to occur again. Especially if homologous recombination is the mechanism, the duplicated unit will still be homologous to neighboring regions, making another duplication of the same region more probable. Indeed, in the Nebulin gene, we have found many multiple tandem duplications at approximately the same time point. Hence, it is possible that most of the vertebrate nebulin repeats were expanded during a very short time period.

The functional implications of the duplications are not easy to determine. The suggested function of Nebulin is for it to act as a ruler determining myofibril length.[11] Therefore, it is possible that the increased length of Nebulin, resulting from duplication events, affects myofibril length in some tissues. Still, the duplicated region in the human SR has so far only been observed in a minor transcript in fetal cells; hence, its functional implications may be small. We investigated the selective pressure on the recently duplicated regions, but we could not find any indication that these regions would be more (or less) conserved than any other part of the protein (data not shown). Thus, although the sequencing quality is too poor for firm conclusions, we can attest to the absence of evidence for positive or negative selection in these regions.

Our study of Nebulin proteins indicates that these tandem duplications are ongoing processes. However, if such duplications involve coding regions of nonrepeating proteins, the resulting gene products are less likely to be functional and would therefore not be retained in the population. On the contrary, if a tandem duplication occurs in a long repeat region, it may have less impact on the overall structure of the protein and hence will not be under strong negative selection. In agreement with this theory, we show that tandem duplications within gene regions are overrepresented in proteins with long domain repeats.

In the literature, there are several examples of tandem duplication polymorphisms that have been related to different human diseases[31–34] in both repeated and nonrepeated proteins. For instance, it has been shown that the duplication of 4 exons containing six cysteine-rich repeats in the low-density lipoprotein receptor produces a functional protein; however, patients with this duplication suffer from familial hypercholesterolemia.[32] In addition, we see that a region duplicated in human Nebulin is associated with copy number polymorphisms. At least 24% of all human copy number variants are associated with segmental duplications.[35] Hence, all segmental duplications are in fact not fixed in humans, but rather correspond to copy number variations in a reference genome. Thus, with the sequencing of several more personal genomes, it will be interesting to see what impact tandem duplication has on shaping our human genome. It is possible that there are individual variations in many other repeat regions similar to the duplications seen for vertebrate Nebulin.

Methods

Data

The gene sequences for orthologs to the human nebulin-containing genes (LASP, Nebulette, NRAP, and Nebulin) from all fully sequenced genomes in Ensembl were downloaded together with exon annotations‡. The data from most genomes are not presented in this article, as many had incomplete coverage of the sequences of the genes of interest. A list of genes and assignments for a selection of species is presented in Table 1. Still, many of the genes presented in the table have unsequenced regions within the genes or are even truncated at some instances.

Nebulin exon assignments

Profiles were created for exons annotated in Ensembl for all mouse and human genes containing nebulin domains using PSI-BLAST,[36] five iterations, and an *e*-value cutoff of 10^{-3}. The resulting profile database was used to search for all three reading frames of nebulin-containing genes in all species using RPSblast[37] and an *e*-value cutoff of 10^{-3}.

To define the NEU within the human Nebulin SR, we split up the longer exons into two or three domains, with boundaries corresponding to the boundary of the single exon units. Another profile for all simple repeat exons was also created. The resulting eight units were used to search for similar NEU outside the region with well-defined 2, 1, 1, 3 exons (Fig. 1). The profiles from the human SR were also used to find SR NEU in other vertebrates.

However, the human SR is too distant for a successful assignment of the *B. floridae* SR region. Hence, new profiles for assigned exons were created for each of the two SRs in *B. floridae* in accordance with Putnam *et al.*[38] These were then used for assignment in this organism and for ordering of NEU in the *B. floridae* SR (i.e., defining which of the NEU should be regarded as the first in the SR). To determine the order, we used the alignments to human Nebulin SR units and optimized the order to get a high sequence similarity, as seen in the similarity matrices in Fig. 2.

‡ ftp://ftp.ensembl.org/

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011

Alignments

All pairwise sequence alignments were performed using the Smith–Waterman alignment tool in the EMBOSS package[39] and default parameters. Multiple alignments were created using Kalign at default settings,[40] except in the case of the multiple sequence alignment in Fig. 4 where a lower gap extension at 0.1 was used. Visualization of alignments for the figures in this article was created with Jalview.[41]

Reciprocal best hits among the human nebulin-containing proteins were identified as pairs of exons that have the highest alignment scores among all exons in the four human paralogs.

Identification of duplication events

For identification of recent duplication events, the DNA sequence of each gene was aligned to itself and visualized using the dotplot program Dotter.[42] In the alignment visualization, highly similar stretches were manually identified. In our summary of duplication events in Fig. 4 and Table 2, only events that involve at least one intron pair with >90% sequence identity are included. These events stand out clearly from the background distribution, with about 60–65% sequence identity between all pairs of introns within one Nebulin gene. We can also conclude that such duplications are independent events in each species, since the conservation of intron sequences between species is much lower than that. In fact, the average identity between corresponding human and mouse introns is 66%, and that between human and platypus is 61%; the best-conserved introns have 81% and 78% identity, respectively.

Some of the events have a different number of exons in the duplicate copies (events 4, 7, and 9 in Fig. 4). This may reflect incomplete duplication of the SR, or exon deletion subsequent to duplication. However, many of these instances overlap with sequencing gaps and could also reflect incomplete sequencing of the concerned exons.

Duplications in the SR regions preceding the split of mammals and other vertebrates, as shown in Fig. 3, were identified as gaps in the multiple alignment of mouse to nonmammalian species. The mouse was selected since it did not have any SR duplications after the mammalian split. The multiple alignment is presented in Supplementary Material.

Miscellaneous methods

Clustering was performed using the PHYLIP package and the neighbor joining algorithm.[43] DNA repeats were identified using the repeatmasker program[23] and a cutoff score of 225. Sequence logos were created using the WebLogo program.[44] Protein domains were assigned with HMMER-3§ and HMM models from Pfam.[18] Segmental duplication data were downloaded from the Human Segmental Duplication Database.[22]

§ http://hmmer.wustl.edu

Acknowledgements

This work was supported by grants from the Swedish Research Council, SSF (the Foundation for Strategic Research). The EU 6th Framework Program is gratefully acknowledged for supporting the EMBRACE project (contract no. LSHG-CT-2004-512092).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.07.011

References

1. Apic, G., Gough, J. & Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325.
2. Vogel, C., Teichmann, S. & Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. *J. Mol. Biol.* **346**, 355–365.
3. Björklund, Å. K., Ekman, D., Light, S., Frey-Skött, J. & Elofsson, A. (2005). Domain rearrangements in protein evolution. *J. Mol. Biol.* **353**, 911–923.
4. Weiner, J., III, Beaussart, F. & Bornberg-Bauer, E. (2006). Domain deletions and substitutions in the modular protein evolution. *FEBS J.* **273**, 2037–2047.
5. Ekman, D., Björklund, Å. K., Frey-Skött, J. & Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of like-orphan domains and other unassigned regions. *J. Mol. Biol.* **348**, 231–243.
6. Björklund, Å. K., Ekman, D. & Elofsson, A. (2006). Expansion of protein domain repeats. *PLoS Comput. Biol.* **2**, e114.
7. Marcotte, E., Pellegrini, M., Yeates, T. O. & Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* **293**, 151–160.
8. Ekman, D., Light, S., Björklund, Å. K. & Elofsson, A. (2006). What properties characterize the hub proteins of the protein–protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol.* **7**, R45.
9. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309**, 1–18.
10. Pfuhl, M., Winder, S. & Pastore, A. (1994). Nebulin, a helical actin binding protein. *EMBO J.* **13**, 1782–1789.
11. McElhinny, A. S., Kazmierski, S. T., Labeit, S. & Gregorio, C. C. (2003). Nebulin: the nebulous, multifunctional giant of striated muscle. *Trends Cardiovasc. Med.* **13**, 195–201.
12. Root, D. & Wang, K. (2001). High-affinity actin-binding nebulin fragments influence the actoS1 complex. *Biochemistry*, **40**, 1171–1186.
13. Donner, K., Sandbacka, M., Lehtokari, V., Wallgren-Pettersson, C. & Pelin, K. (2004). Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts. *Eur. J. Hum. Genet.* **12**, 744–751. doi:10.1038/sj.ejhg.5201242.

14. Moncman, C. & Wang, K. (2002). Targeted disruption of Nebulette protein expression alters cardiac myofibril assembly and function. *Exp. Cell Res.* **273**, 204–218. doi:10.1006/excr.2001.5423.
15. Li, B., Zhuang, L. & Trueb, B. (2004). Zyxin interacts with the SH3 domains of the cytoskeletal proteins LIM-Nebulette and Lasp-1. *J. Biol. Chem.* **279**, 20401–20410.
16. Zieseniss, A., Terasaki, A. G. & Gregorio, C. C. (2008). Lasp-2 expression, localization, and ligand interactions: a new Z-disc scaffolding protein. *Cell Motil. Cytoskeleton*, **65**, 59–72.
17. Hanashima, A., Kubokawa, K. & Kimura, S. (2009). Characterization of amphioxus nebulin and its similarity to human nebulin. *J. Exp. Biol.* **212**, 668–672. doi:10.1242/jeb.022681.
18. Sonnhammer, E., Eddy, S. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins Struct. Funct. Genet.* **28**, 405–420.
19. Kontrogianni-Konstantopoulos, A., Ackermann, M., Bowman, A., Yap, S. & Bloch, R. (2009). Muscle giants: molecular scaffolds in sarcomerogenesis. *Physiol. Rev.* **89**, 1217–1267.
20. Zhang, J., Weisberg, A. & Horowits, R. (1998). Expression and purification of large nebulin fragments and their interaction with actin. *Biophys J.* **74**, 349–359.
21. Rohl, C., Strauss, C., Misura, K. & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93.
22. Bailey, J., Yavor, A., Massa, H., Trask, B. & Eichler, E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017. doi:10.1101/gr.187101.
23. Smit, A., Hubley, R. & Green, P. (1996–2004). Repeatmasker open-3.0. http://www.repeatmasker.org.
24. Perry, G., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. *et al.* (2008). The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* **82**, 685–695. doi:10.1016/j.ajhg.2007.12.010.
25. Tuzun, E., Sharp, A., Bailey, J., Kaul, R., Morrison, V., Pertz, L. *et al.* (2005). Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732. doi:10.1038/ng1562.
26. Conrad, D., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712. doi:10.1038/nature08516.
27. Pappas, C., Bhattacharya, N., Cooper, J. & Gregorio, C. (2008). Nebulin interacts with CapZ and regulates thin filament architecture within the Z-disc. *Mol. Biol. Cell*, **5**, 1837–1847.
28. Witt, C., Burkart, C., Labeit, D., McNabb, M., Wu, Y., Granzier, H. & Labeit, S. (2006). Nebulin regulates thin filament length, contractility, and Z-disk structure *in vivo*. *EMBO J.* **25**, 3843–3855.
29. Koszul, R. & Fischer, G. (2009). A prominent role for segmental duplications in modeling eukaryotic genomes. *C. R. Biol.* **332**, 254–266.
30. Bailey, J., Liu, G. & Eichler, E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834.
31. Strout, M. P., Marcucci, G., Bloomfield, C. D. & Caligiuri, M. A. (1998). The partial tandem duplication of ALL1 (MLL) is consistently generated by Alu-mediated homologous recombination in acute myeloid leukemia. *Proc. Natl Acad. Sci.* **95**, 2390.
32. Patel, D. (1998). Analysis of two duplications of the LDL receptor gene affecting intracellular transport, catabolism, and surface binding of the LDL receptor. *J. Lipid Res.* **39**, 1466.
33. Heikkinen, J., Toppinen, T., Yeowell, H., Krieg, T., Steinmann, B., Kivirikko, K. & Myllylä, R. (1997). Duplication of seven exons in the lysyl hydroxylase gene is associated with longer forms of a repetitive sequence within the gene and is a common cause for the type VI variant of Ehlers–Danlos syndrome. *Am. J. Hum. Genet.* **60**, 48.
34. Fukao, T., Zhang, G., Rolland, M. O., Zabot, M. T., Guffon, N., Aoki, Y. & Kondo, N. (2007). Identification of an Alu-mediated tandem duplication of exons 8 and 9 in a patient with mitochondrial acetoacetyl-CoA thiolase (T2) deficiency. *Mol. Genet. Metab.* **92**, 375–378.
35. Redon, R., Ishikawa, S., Fitch, K., Feuk, L., Perry, G., Andrews, T. *et al.* (2006). Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
36. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
37. Marchler-Bauer, A., Panchenko, A., Shoemaker, B., Thiessen, P., Geer, L. & Bryant, S. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**, 281.
38. Putnam, N., Butts, T., Ferrier, D., Furlong, R., Hellsten, U., Kawashima, T. *et al.* (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.
39. Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277.
40. Lassmann, T., Frings, O. & Sonnhammer, E. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res.* **37**, 858–865. doi:10.1093/nar/gkn1006.
41. Clamp, M., Cuff, J., Searle, S. & Barton, G. (2004). The Jalview Java alignment editor. *Bioinformatics*, **20**, 426.
42. Sonnhammer, E. & Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, 1–2.
43. Felsenstein, J. (2004). PHYLIP (Phylogeny Inference Package) Version 3.6 Department of Genome Sciences, University of Washington, Seattle, WA; distributed by the author.
44. Crooks, G., Hon, G., Chandonia, J. & Brenner, S. (2004). WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188.

Please cite this article as: Björklund, Å. K., *et al.*, Nebulin: A Study of Protein Repeat Evolution, *J. Mol. Biol.* (2010), doi:10.1016/j.jmb.2010.07.011