Correspondence

The DNA-binding region of RAG1 is not a homeodomain

Sharmila Banerjee-Basu and Andreas D Baxevanis

Address: Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892-4470, USA.

Correspondence: Andreas D Baxevanis. E-mail: andy@nhgri.nih.gov

Published: 25 July 2002

Genome ***Biology*** 2002, **3(8)**:interactions1004.1–1004.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/8/interactions/1004

One of the goals of functional annotation is to catalog information that would be of value in guiding experimental design and analysis. Even in cases for which sequence similarity can be detected reliably, however, functional annotations found in public databases are often incorrect [1,2]. Here, we discuss a case in which a functional assignment was made to RAG1, a protein catalogued as a homeodomain protein in the Online Mendelian Inheritace in Man (OMIM) database [3]. A more in-depth bioinformatic analysis shows this assignment to be incorrect. The known biochemical functions of RAG1 are as an integrase and recombinase [4], functions that are not consistent with those of other homeodomain proteins [5].

Comparison of RAG1 with homeodomains

The homeodomain is a DNA-binding domain found in many eukaryotic transcription factors and is characterized by a highly stringent sequence signature [6,7]. Structural studies on homeodomain family members have revealed that these proteins contain almost superimposable structures, all consisting of a three-helical bundle with an amino-terminal extension and all exhibiting a similar mode of DNA binding [8-10]. Several positions within the homeodomain region that are involved in DNA recognition or stabilization of the structure are conserved across species.

The DNA-binding region of RAG1 is also highly conserved (Figure 1a). This region shows 20% sequence identity with the homeodomain of the Engrailed protein. The DNA-binding domain of RAG1 could not, however, be aligned to a dataset of 129 human homeodomain sequences [11], as RAG1 lacks the evolutionarily conserved amino-acid residues that define the homeodomain family. Also, the sequence motif in the homeodomain DNA-recognition helix (48-WF-x-N-x-R-53, where x is any amino acid) is in fact absent from the RAG1 DNA-binding region. Experimental evidence that RAG1 does not belong to the homeodomain family comes from the observation that a mutant RAG1 protein containing a conserved homeodomain motif failed to bind DNA and is non-functional in *in vitro* recombination assays [12].

Comparison of RAG1 with Tc3 transposase and Hin invertase

Caenorhabditis elegans Tc3 is a member of the Tc1/mariner family of transposable elements found in species ranging from fungi to humans [13]. The X-ray structure of Tc3 transposase revealed that the specific DNA-binding region contains three α helices comprising the helix-turn-helix motif [14]. Sequence alignments between the DNA-binding regions of RAG1 and Tc3 shows patches of high sequence conservation, especially in the amino-terminal region. The lack of the DNA-contacting residues of Tc3 from RAG1 indicates significant divergence in the RAG1 DNA-recognition site, however. Automated fold prediction using the University of California, Los Angeles and the Department of Energy cooperative (UCLA-DOE) Fold Recognition Server [15] identified the DNA-binding domain of Tc3 transposase of *C. elegans* as the candidate whose fold is most likely to represent the RAG1 family.

Hin recombinase belongs to a family of bacterial DNA invertases that catalyze a site-specific recombination reaction. The Hin DNA-binding domain shares distinct sequence similarities with RAG1, and there is a striking similarity between the Hin recognition sequence and RAG1-nonamer site. At the sequence level, the invariant 138-GGRPR-142 motif in the amino-terminal arm of Hin DNA-binding domain is conserved in RAG1. This motif is positioned in the minor groove of the DNA-recognition sequence and provides critical DNA contacts. The sequence similarity between RAG1 and Hin recombinase is extended through the DNA-binding region, with a total of

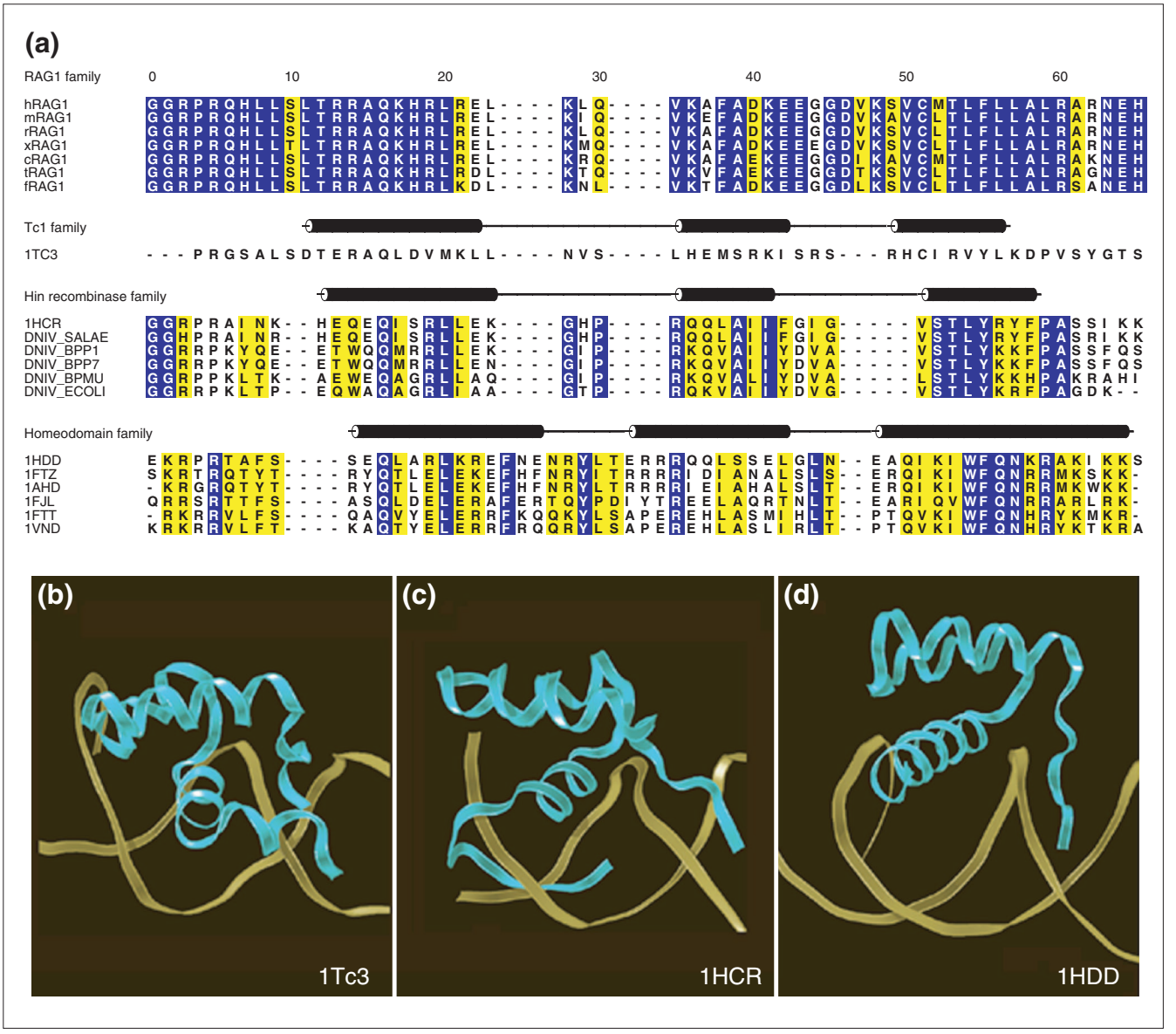

Figure 1
RAG1 family members and related DNA-binding proteins containing helix-turn-helix motifs. **(a)** Multiple sequence alignment of the DNA-binding regions. The alignment was made using CLUSTAL W [20] and manually refined to reflect the secondary structural elements. Homologous DNA-binding regions from RAG1 family members, Tc3 transposase, the Hin family of recombinases, and selected homeodomain proteins are shown in the single-letter amino-acid code. For the RAG1 family the sequences are from human (hRAG1), mouse (mRAG1), rabbit (rRAG1), *Xenopus* (xRAG1), chicken (cRAG1), trout (tRAG1), and *Fugu* (fRAG1), respectively. For the Tc1 family, the sequence is from *Caenorhabditis elegans* Tc3 transposase (1TC3). For the Hin recombinase family, the sequences are from *Salmonella typhimurium* (1HCR), *Salmonella abortus-equi* (DNIV_SALAE), bacteriophage P1 (DNIV_BPP1), bacteriophage P7 (DNIV_BPP7), bacteriophage Mu (DNIV_BPMU), and *Escherichia coli* (DNIV_ECOLI). For the homeodomain family, the sequences are from *Drosophila* Engrailed (1HDD), *Drosophila* Fushi tarazu (1FTZ), *Drosophila* antennapedia (1AHD), *Drosophila* Paired (1FJL), rat Thyroid transcription factor 1 (1FTT), and *Drosophila* NK-2 (1VND). Amino-acid residues showing absolute identity among these proteins are shown with a blue background; those with conservative substitution are shown with a yellow background. The positions of the three α helices defined in the X-ray structures of *C. elegans* Tc3 transposase, *Salmonella* Hin recombinase, and *Drosophila* engrailed homeodomain are schematically represented above respective sequences. ALSCRIPT [21] was used to format the alignment. **(b-d)** Ribbon diagrams of selected helix-turn-helix motif containing protein domains bound to target DNA. The α-carbon backbone of the protein is depicted in blue and the DNA as a yellow ribbon, respectively. Helices and loop regions are as defined in (a). The ribbon diagram was generated using the Visual Molecular Dynamics program [22] with the atomic coordinates from (b) *C. elegans* Tc3 transposase (PDB [23] entry 1Tc3), (c) *Salmonella* Hin recombinase (PDB entry 1HCR), and (d) *Drosophila* engrailed homeodomain (PDB entry 1HDD).

13 residues absolutely conserved in this region. Structural conservation of the DNA-binding domain of RAG1 and Hin recombinase is illustrated by the observation that a RAG1 hybrid protein containing the homologous DNA-binding region of Hin recombinase is functional in *in vitro* recombination assays [12].

The helix-turn-helix motif

In the DNA-binding domains of Hin recombinase, Tc3 transposase, and Engrailed, the first and the second helices lie almost anti-parallel to each other, with a turn between the second and the third helices. In all cases, the recognition helix fits into the major groove of the DNA. Although the essential features of the helix-turn-helix motifs are very similar, these proteins do not all dock on the DNA in the same fashion (Figure 1b-d).

In the X-ray structure of the Engrailed homeodomain-DNA complex, several residues in the exposed hydrophilic face of helix 3 establish specific contacts with the last four base pairs of the recognition sequence, whereas the residues in the amino-terminal arm of the protein contact the first two base pairs of the recognition sequence. Compared to Tc3 transposase and Hin recombinase, the helices and the loops are longer in the homeodomain structure; only helix 3 is inserted in the major groove and the residues in the center of this relatively longer helix provide DNA contacts.

In the Hin-recombinase structure, the α-helical core, along with extensions at both the amino and carboxyl termini, participate in DNA recognition. The eight-residue carboxy-terminal tail of Hin recombinase is inserted in the minor groove of the DNA-recognition site. Wrap-around of the DNA-binding site by the carboxy-terminal extension has not been observed in Tc3 or homeodomain structures. In contrast to the other structures, both the second and third helices of Tc3 transposase participate in DNA recognition by binding to the major groove. The six residues preceding the first helix in Tc3 adopt a conformation different from that seen in the longer amino terminus of the Hin recombinase and the Engrailed homeodomain.

Genomic perspective

The ability of the RAG1 proteins to catalyze both the formation of hybrid joints and transposition highlight the similarities between the mechanism of site-specific rearrangement by V(D)J recombination and certain transposition/retroviral integration reactions. The occurrence of RAG proteins in jawed vertebrates and conservation of domain architecture and function from prokaryotes suggest that the RAG1 proteins might have been horizontally transferred into the eukaryotic genome by a transposon.

The question that may be posed here is what the relevance of the current observation is, and whether the functional mis-assignment is of great importance. During vertebrate lymphocyte development, RAG1 mediates the somatic assembly of antigen receptors, which involves DNA-bond breakage and strand-transfer reactions, reminiscent of transposition reactions in bacteria. Homeodomain proteins play a fundamental role in diverse cellular processes by transcriptional regulation of downstream-target genes. RAG1 has been identified only in jawed vertebrates, whereas homeodomain proteins are highly conserved from yeast to human. The evolution and biological functions of RAG1 and homeodomain proteins are markedly different, and one cannot substitute for the other.

Unfortunately, with the initial misclassification by Spanopoulou *et al.* [12] has come experimental interpretation in the context of RAG1 being a homeodomain. Specifically, Villa *et al.* [16] have interpreted the biochemical effects of mutations leading to Omenn Syndrome as having to do with changes in homeodomain structure, despite statements implicating the observed defects with low degrees of V(D)J recombination. In addition, Aidinis *et al.* [17] proposed models of interaction of a RAG1 'homeodomain' with the chromatin proteins HMG1 and HMG2. In the study by Aidinis *et al.* [17], the experiments were designed under the assumption that RAG1 was a homeodomain, leading to incorrect extension of the interpretation of results to the involvement of a homeodomain structure in V(D)J recombination. The model proposed by this group has therefore been made in the wrong biological context.

The incorrect assignment of RAG1 as a homeodomain has colored the interpretation of experimental results. This is emblematic of the larger problem that annotation-error propagation plays in incorrectly guiding experimental discovery. Often, there may be little or no similarity between a sequence of interest and those in the public databases, meaning that it would be very difficult (if not impossible) to determine any degree of relatedness on the basis of sequence alone. Even in cases where homology can be detected reliably, the annotations currently found in the public databases are often incorrect. The considerable effect of processes such as alternative splicing [18] and the ability of proteins to perform markedly different functions depending on their cellular localization and compartmentalization [19], coupled with the number of annotation errors currently in the public databases, all help to re-emphasize the importance of database curation and experimental validation in maintaining the purity and utility of these public resources.

References

1. Brenner SE: **Errors in genome annotation.** *Trends Genet* 1999, **15:**132-133.
2. Baxevanis AD: **Making the best use of publicly-available bioinformatics resources: keeping biology in mind.** *Nature Genetics* 2002, *in press*.
3. **Online Mendelian Inheritance in Man** [http://www.ncbi.nlm.nih.gov/omim/]
4. Sadofsky MJ: **The RAG proteins in V(D)J recombination: more than just a nuclease.** *Nucleic Acids Res* 2001, **29:**1399-1409.
5. Gehring WJ, Affolter M, Burglin T: **Homeodomain proteins**. *Annu Rev Biochem* 1994, **63**:487-526.

6. Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M, Otting G, Wuthrich K: **Homeodomain-DNA recognition.** *Cell* 1994, **78:**211-223.
7. Laughon A: **DNA binding specificity of homeodomains.** *Biochemistry* 1991, **30:**11357-11367.
8. Wolberger C, Vershon AK, Liu B, Johnson AD, Pabo CO: **Crystal structure of a MATα2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions.** *Cell* 1991, **67:**517-528.
9. Kissinger CR, Liu BS, Martin-Blanco E, Kornberg TB, Pabo CO: **Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions.** *Cell* 1990, **63:**579-590.
10. Gruschus JM, Tsao DH, Wang LH, Nirenberg M, Ferretti JA: **Interactions of the vnd/NK-2 homeodomain with DNA by nuclear magnetic resonance spectroscopy: basis of binding specificity**. *Biochemistry* 1997, **36**:5372-5380.
11. Banerjee-Basu S, Baxevanis AD: **Molecular evolution of the homeodomain family of transcription factors.** *Nucleic Acids Res* 2001, **29:**3258-3269.
12. Spanopoulou E, Zaitseva F, Wang FH, Santagata S, Baltimore D, Panayotou G: **The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination.** *Cell* 1996, **87:**263-276.
13. Doak TG, Doerder FP, Jahn CL, Herrick G: **A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common 'D35E' motif.** *Proc Natl Acad Sci USA* 1994, **91:**942-946.
14. van Pouderoyen G, Ketting RF, Perrakis A, Plasterk RH, Sixma TK: **Crystal structure of the specific DNA-binding domain of Tc3 transposase of *C. elegans* in complex with transposon DNA.** *EMBO J* 1997, **16:**6044-6054.
15. **UCLA-DOE fold server** [http://fold.doe-mbi.ucla.edu/]
16. Villa A, Santagata S, Bozzi F, Giliani S, Frattini A, Imberti L, Gatta LB, Ochs HD, Schwarz K, Notarangelo LD, *et al.*: **Partial V(D)J recombination activity leads to Omenn syndrome.** *Cell* 1998, **93:**885-896.
17. Aidinis V, Bonaldi T, Beltrame M, Santagata S, Bianchi ME, Spanopoulou E: **The RAG1 homeodomain recruits HMG1 and HMG2 to facilitate recombination signal sequence binding and to enhance the intrinsic DNA-bending activity of RAG1-RAG2.** *Mol Cell Biol* 1999, **19:**6532-6542.
18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
19. Jeffery CJ: **Moonlighting proteins**. *Trends Biochem Sci* 1999, **24**:8-11.
20. **ClustalW** [http://www.ebi.ac.uk/clustalw/]
21. Barton, GJ: **ALSCRIPT: a tool to format multiple sequence alignments.** Protein Eng 1993, **6:**37-40.
22. **Visual Molecular Dynamics** [http://www.ks.uiuc.edu/Research/vmd]
23. **Protein Data Bank (PDB)** [http://www.rcsb.org/pdb/]