New insights into the evolution of C-1 family cadherins

M A Hasenahuer, V H Casco and M F Izaguirre


E-mail: fizaguirre@bioingenieria.edu.ar

Abstract. Specific cell-cell adhesion and intracellular communication are key processes in multicellular animals, and members of the cadherin superfamily are essential players in these processes. Evolutionary origin of the modern C-1 family in metazoan has been established based on bioinformatic tools using Neighbor-Joining (NJ) algorithm and the EC1 first cadherin-repeat sequence—a very narrow segment of the protein full-length but responsible of adhesive recognition—and their cytoplasmic domain. Searching for insights that reveal organizational changes in cadherin superfamily C-1 members, a phylogenetic analysis of full extracellular (EC) and cytoplasmic domains (CD) was performed, using Maximum Likelihood (ML) phylogenetic algorithm. We compiled literature data and performed an in-depth phylogenetic analysis of 145 members of this superfamily from 8 species, covering several representative branches within metazoan evolution. We analyzed the amino acids sequence homology between extracellular and cytoplasmic domains, and we reviewed protein structural data architecture. Analysing EC full-length with the Maximum Likelihood approach we could reclassify the solitary CDH13 group, as keeping within the classical type I cadherins.

1. Introduction

Cadherins are a large family of \( \text{Ca}^{2+} \)-dependent cell surface adhesion glycoproteins, temporal and tissular differentially expressed in many organisms. They are essential for animal development and tissues integrity [1, 2, 3, 4, 5]. Their deficiency has been associated with tumor progression and metastasis in mammals [6, 7].

Cadherins have been identified in many organisms from unicellular choanoflagellates to invertebrates and all classes of vertebrates [1, 2, 8]. Day-to-day new cadherin-like molecules are included in the superfamily, and actually, a new phylogenetic classification organize them in a Cadherin Major Branch (CMB) that includes C-1 and C-2 Families, and others; Cadherin-related Major Branch (CrMB) that includes Cr-1a (protocadherins), Cr-1b, Cr-2 (CDHR) and Cr-3; and Curiosa that includes Solitary, w/o ECs and Primitive [8].

In their mature state, many of them have three segments: extracellular (EC), trans-membrane (TM) and cytoplasmic (CD) domains. EC is involved in mainly homophilic recognition and adhesion mechanisms and is constituted by a variable amount of cadherin-repeats (ECs) (Pfam: PF00028), each one of them has a \( \beta \)-sandwich Ig-like folding of ~110 residues that contains conserved Ca\(^{2+}\)-coordinating regions [9]. Each EC is enumerated from the outermost N-terminal EC1 to the closest to membrane ECn. CD is more variable among cadherin subfamilies [8], and binds to different intracellular proteins to give functional connection to the cytoskeleton [9].

Many EC1-dependent adhesion models have been proposed based on known structures. First, through a strand-swapping interface composed by a tryptophan in position 2 (Trp2) and many other
residues shaping a hydrophobic pocket where a Trp2 from another cis (from the same cell surface) or trans (from an adjacent cell) cadherin inserts, and vice versa [10, 11, 12]. Another cis dimerization interface has been postulated involving a coupling between a convex surface at the bottom of EC2 of a monomer and a concave surface at the EC1 of another one [9], while allowing them to perform the trans strand-swapping. Furthermore, a particular X-shaped dimerization interface has been observed in CDH1 and CDH13, near the EC1-EC2 calcium-binding sites [13, 14].

Owing to the fact that cadherin superfamily members differ greatly in protein length and domain organization —some of them have two EC repeats and less than 1000 AA, while others have over 4000 AA and 34 EC repeats—a reliable representation of the evolutionary relationships between the diverse cadherin molecules is therefore not feasible using the entire protein sequences. Instead, over the past years several researchers have used for phylogenetic studies on cadherin superfamily members either the first EC (EC1) repeat, which is responsible for adhesive interactions in classic cadherins [10, 11, 12], or the cytoplasmic domain [8]. In contrast, thinking that the analysis of EC full-length could reveal phylogenetic information not provided by only analyzing EC1, in the present work, a phylogenetic analysis of C-1 family full extracellular and cytoplasmic domains was performed. C-1 family comprises the ‘classical’ type-I cadherins, the ‘atypical’ type-II cadherins, the ‘desmosomal’ desmogleins and desmocollins, the ‘7D’-cadherins and the ‘isolated’ CDH13 and CDH26 cadherins (Table 1) [8, 15]. Because of our limited computational capacity, only the C-1 cadherins were analyzed.

<table>
<thead>
<tr>
<th>Table 1. C-1 cadherin family members (Modified from [8]).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subfamily</strong></td>
</tr>
<tr>
<td><strong>Type I</strong></td>
</tr>
<tr>
<td><strong>Type II</strong></td>
</tr>
<tr>
<td><strong>Desmocollins</strong></td>
</tr>
<tr>
<td><strong>Desmogleins</strong></td>
</tr>
<tr>
<td><strong>7D</strong></td>
</tr>
</tbody>
</table>
2. Materials and methods

2.1. Protein data set preparing and multiple sequence alignment

Homologous protein sequences of C-1 cadherin members of: human (Hs), mouse (Mm), rat (Rn), cow (Bt), chicken (Gg), zebrafish (Dr), african clawed frogs (Xl) and (Xt), were retrieved from UniProtKB [http://www.uniprot.org/] in FASTA format. Search was done by keyword and BLASTP.

Because 7D cadherins have seven EC repeats —instead of five like all other C-1 members—, it was necessary to decide which ECs to include. CDH17’s EC3-EC7 segment was found to be more similar to the EC1-EC5 of classical cadherins [16], so only these five repeats were considered from de 7D members. For some cadherins in a given species, only unreviewed sequences and others named putative uncharacterized protein were found, but showing a high similarity to the reviewed suspected orthologs from other species. With the aim of include as many sequences as possible, they were taken into account only after a careful visual comparative inspection against the well characterized possible orthologs.

Inspection and a posterior edition to separately analyze EC and CD were carried out using MEGA5 software [17]. To compare only homologous characters, as required for phylogenetical analysis, incomplete and spurious sequences were not included, leaving a total of 145 utilizable ones.

Multiple sequence alignment of EC and CD data sets were carried out in MUSCLE web server [http://www.ebi.ac.uk/Tools/msa/muscle/], using parameters by default.

Because of the sequence variability between CDs from different C1-subfamilies a low quality alignment was obtained when trying to align all CDs together. For this reason, they were separated and aligned according to the four subfamilies (i.e. type I CDs, type II CDs, DSG CDs and DSC CDs). Owing to CDH13 lacks CD —is attached to the membrane by a glycosylphosphatidylinositol (GPI) anchor [18]—, CDH26 does not share significant CD similarity with other members and 7Ds have a very short CD, their CD trees were not constructed.

All five alignments were manually inspected and corrected where necessary, bearing in mind information about structure and domain conservation. Ambiguous gapped regions were wiped out.

2.2. Testing for evolutionary model fitting

Once alignments were ready, each one was transformed into PHYLIP format with the ReadSeq format conversion tool [http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html].

Each PHYLIP formatted alignment was tested with ProtTest in order to find the most suitable aminoacid substitution model that best represents their evolution [19]. Only matrices applicable to the data sets were chosen to test for best fitting: JTT, LG, DCMut, Dayhoff, WAG, Blosum62, and VT; the remaining, for mitochondrial, retroviral and reverse transcriptase proteins were not taken into account [20, 21]. Estimations of +G (gamma function shape), +F (aminoacids frequency) and +I (proportion of invariant sites) parameters with ten categories for gamma distribution (each category corresponds to a different rate of evolution) were set too [20, 21].

2.3. Phylogenetic reconstruction and statistical validation

Once best fitting matrices and suggested parameters (+G, +F and/or +I) to take into account were found, they were used as input settings to launch each phylogenetic tree inference by Maximum Likelihood (ML) method in PhyML3.0 software [22]. For the EC-based tree, statistical validation was done with 500 non-parametric bootstrap replicates, whereas for CD-based trees with 1,000 replicates.

Owing to the computational burden [23], the replicates number was selected for each data set size —145 EC-sequenceset, 500 replicates and four sets of no more than 65 CD-sequences each one, 1000 replicates—. BioNJ was chosen as the calculation method of the starting trees to being optimized. Twenty gamma categories and both, NNI (nearest neighbor interchange) and SPR (subtree pruning
and regrafting) optimizing algorithms were selected. Twenty is considered the “top reasonable” number by PhyML developers. SPR and NNI are two different ways to optimize a tree topology, so we selected the BEST option which includes the execution of both methods, and returns the best solution (with the highest likelihood) among the two [21]. ML with bootstrapping validation is statistically robust but is a computationally-demanding method [23]. PhyML3.0 allows using parallel computing to perform bootstrap analysis, therefore these calculations could be executed through MPI (Message Passing Interface), using 10 parallel processors for each tree in a cluster of computers.

7Ds cadherins were considered as outgroup for EC based tree. This is because some authors propose they had a common ancestor to all other C-1 cadherins, but have diverged from them by acquiring two extra ECs after duplication events, becoming a separate group [24].

3. Results and discussion

3.1. EC and CD best-fitting aminoacid substitution models

Best fitting evolution models, according to a consensus between AICs (Akaike Information Criterion) and BICs (Bayesian Information Criterion) frameworks were: LG+I+G for EC and JTT+G for all CDs (Table 2) [20].

Table 2: Weights (probabilities of being the best) of the different models under the different frameworks for each data set. “(1)” represents the best models for each of the 7 frameworks. “*” depicts the chosen model when different frameworks disagree about which is the best. Only rows of models were at least one (1) appeared are shown.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Model</th>
<th>AIC</th>
<th>AICc-1</th>
<th>AICc-2</th>
<th>AICc-3</th>
<th>BIC-1</th>
<th>BIC-2</th>
<th>BIC-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC</td>
<td>LG+I+G</td>
<td>1.00(1)</td>
<td>1.00(1)</td>
<td>1.00(1)</td>
<td>1.00(1)</td>
<td>1.00(1)</td>
<td>1.00(1)</td>
<td>1.00(1)</td>
</tr>
<tr>
<td>CD typeI</td>
<td>JTT+G</td>
<td>0.73(1)</td>
<td>0.95(1)</td>
<td>0.99(1)</td>
<td>0.76(1)</td>
<td>0.92(1)</td>
<td>0.92(1)</td>
<td>0.97(1)</td>
</tr>
<tr>
<td>CD typeII</td>
<td>JTT+G</td>
<td>0.61(1)</td>
<td>0.00(29)</td>
<td>1.00(1)</td>
<td>0.63(1)</td>
<td>0.87(1)</td>
<td>0.89(1)</td>
<td>0.97(1)</td>
</tr>
<tr>
<td>CD DSG</td>
<td>JTT+G+F</td>
<td>0.64(1)</td>
<td>0.01(3)</td>
<td>0.05(3)</td>
<td>0.53(1)</td>
<td>0.00(4)</td>
<td>0.00(4)</td>
<td>0.00(5)</td>
</tr>
<tr>
<td></td>
<td>JTT+G*</td>
<td>0.09(3)</td>
<td>0.75(1)</td>
<td>0.71(1)</td>
<td>0.21(2)</td>
<td>0.95(1)</td>
<td>0.95(1)</td>
<td>0.97(1)</td>
</tr>
<tr>
<td>CD DSC</td>
<td>JTT+I+G</td>
<td>0.55(1)</td>
<td>0.65(1)</td>
<td>0.66(1)</td>
<td>0.69(1)</td>
<td>0.36(2)</td>
<td>0.34(2)</td>
<td>0.22(2)</td>
</tr>
<tr>
<td></td>
<td>JTT+G*</td>
<td>0.20(2)</td>
<td>0.35(2)</td>
<td>0.34(2)</td>
<td>0.27(2)</td>
<td>0.64(1)</td>
<td>0.66(1)</td>
<td>0.78(1)</td>
</tr>
</tbody>
</table>

AIC: Akaike Information Criterion framework.
AICc-x: Second-Order Akaike framework.
BIC-x: Bayesian Information Criterion framework.
AICc/BIC-1: sample size as: number of sites in the alignment.
AICc/BIC-2: sample size as: sum of position's Shannon Entropy over the whole alignment.
AICc/BIC-3: sample size as: align. length x num sequences x averaged (0-1)Sh. Entropy.

As can be seen, EC and CDs clearly follow different substitution matrices (LG vs. JTT), which means that relative rates of replacement from one aminoacid to another are not equal between EC and CD. We also carried out alignments, substitution model test and phylogenetic inference of the complete EC of each subfamily separately (data not shown). Type I, DSC, DSG best substitution matrices were LG, only type II fitted to a JTT matrix. It suggests that ECs and CDs have evolved by separate.

3.2. EC- and CD-based trees

EC-based phylogenetic tree shows a great divergence of C-1 family members, that spreads branches in three main groups: type II cadherins (cyan branches); type I cadherins, desmocollins, desmogleins and CDH13 (fuchsia branches); and CDH16 and CDH17 (blue branches) (Figure 1). In Figure 2, we present each obtained tree in circular representation with corresponding bootstrap support values over
each internal branch. Considerable confidence can be given to branches or groups supported by more than 70-75%, less supported branches should be treated with caution [23], they are usually collapsed and seen as unresolved politomies. In EC tree 75% represents 375 bootstraps and in CD trees 750 bootstraps.

Although when many columns in the alignments had to be removed to left only the homolog sites, each CD tree demonstrates—through of their shorter distances—that CDs are more conserved in each group of orthologs than ECs (Figures 1 & 2). Our tree clearly shows that ECs are much more conserved among type II cadherins, compared to the other members (Figure 1, see branch lengths).

DSG1 was excluded from EC analysis because it has a shorter EC5 compared to the other C-1 members. Owing to the DSG1’s EC5 shortness, alignment algorithm introduces many continuous gaps to fill this place. In phylogenetic analysis only homologous regions are compared, and our goal was to compare ECs as complete as possible, so we decided not to eliminate the fragment of the alignment where DSG1 was full of gaps, instead we did not include DSG1 sequences in the tree reconstruction. The CD-based phylogenetic analysis shows that DSG1 has a common ancestor with DSG4 (Figures 1 & 2c), confirming others investigations [8].

Interestingly, analyzing EC full-length we could find that CDH13 belongs to type I cadherins (Figure 2a), sharing a common ancestor with CDH2 and CDH4. It seems like there was a duplication event that gave rise to CDH13 in one side and an ancestor on the other hand that, after other duplication event, originated CDH2 and CDH4. In contrast, EC1-based analysis had placed CDH13 together with CDH16 and CDH17 [8], probably because their EC1 differs from the other C-1 cadherins. Our ECn, (n=1…5)-based analysis shows CDH13 as a Type I-member because the remaining EC repeats are more similar to Type I than 7D members. Branches that lead to CDH2/CDH4 and CDH13 members since the duplication event happened, have both 500 bootstraps support, which represents 100% of confidence that such branches are correct.

It is known that CDH13 has an Ile2 instead Trp2 and is thought to dimerize in a weaker way through an X-dimer interface. CDH1 also has been shown to follow the X-dimer way, but only when the strand-swap interface is altered in some way [13]. This evidence suggests that CDH13 separated from CDH2 and CDH4 by losing his CD and changing some of the aminoacids involved in strand-swapping, evolving to an X-dimer weaker adhesion.

Although when it was thought that CDH13 (truncated (T)-cadherin or heart (H)-cadherin) was an exclusive protein of Tetrapoda (“four-footed”) [8], from our phylogenetic analysis we could find a CDH13 protein sequence of zebrafish (UniProtKBID: Q1L8D6), which belongs to Osteichtyes (“bony fishes”).

T-cadherin is a unique cadherin cell adhesion molecule that is anchored to the cell surface membrane through a glycosyl phosphatidyl inositol (GPI) moiety. The cytoplasmic domain, which T-cadherin lacks, is believed to be critical for homophilic binding through interaction with submembrane cytoskeletal proteins. However, the T-cadherin EC domain has been well conserved through evolution in vertebrates, suggesting that T-cadherin may have biological significance in higher animals. Consistent with this hypothesis, recent studies have thrown light on the relevance of T-cadherin in the fields of oncology, neurology, respirology and cardiovascular physiology [25]

Zebrafish CDH13 still keeps as an unreviewed novel protein similar to CDH13 in UniProtKB, but shares 64%, 62%, 62%, 60%, 60% and 60% identity with chicken, mouse, rat, frog, bovine and human CDH13s, respectively (percentages obtained by doing BLASTP alignments of zebrafish CDH13 sequence against UniProtKB). These identities are good signals of protein sequence and structure homology [26]. Furthermore, in our phylogenetic analysis zebrafish CDH13 is positioned close to the others CDH13s, as expected according to the tree of species evolution. These evidences suggest that CDH13 is not a tetrapoda-exclusive cadherin because it is present in osteichtyes too. Further genomes sequencing of species in this lineage and other ancestral ones will provide more precisely information about CDH13 emergence.
The appearance of desmosomal cadherins was thought to happen only in mammals [8]. But, there is DSC2-like in zebrafish (Osteichtyes) and DSC3 in frogs (Lissamphibias). No DSG was found for non-mammal species.

On the other hand, the putative uncharacterized proteins (depicted with ‘*’ on each OTU in Figure 2), grouped as was expected, confirming their orthology with the well identified sequences.

**Figure 1.** Unrooted tree from the C-1 cadherin family. Branches were colored according to the grouping pattern. Scale bar represents 0.1 expected aminoacid residue substitutions per site.
Figure 2. Rooted circular trees of C-1 cadherin family: EC-based (a), Type I and II CD-based (b), DSC and DSG CD-based (c). Bootstrap values are depicted over each internal branch. Asterisks (‘*’) denote the putative uncharacterized proteins. Scale bar represents 0.1 expected aminoacid residue substitutions per site.

Despite the fact that Expresso3DCoffee [24] is strongly recommended to improve multiple alignments accuracy, because uses available structural information of close homolog proteins (60% identity) [26], it only can align up to 100 sequences. Additionally, EC full structures are only available for some type I members and as we are working with different subfamilies, it is likely that a close homolog structure could not represent all sequences. Therefore, we used the MUSCLE web server, which is recommended for aligning 100-500 sequences approximately globally alignable [27].

NJ algorithm employed to construct EC1-based tree [8] is a clustering method based on pair-wise distance matrices that infers the phylogenetic relations and quickly builds trees with a considerable
amount of sequences. On the other hand, the Maximum Likelihood algorithm is based on an optimally search criterion and examines different tree topologies for a given taxa, searching for the tree that optimizes such criterion. It considers that a tree gave rise to the observed alignment, given a specific evolutionary model. This allows the possibility to compare the relative support for different phylogenetic trees in a statistical framework. Although when ML is statistically a good method doubt to the amount of possible trees given for a certain amount of taxons, ML becomes a computationally-demanding method.

4. Conclusions
We have inferred phylogenetic trees using the full-length EC of C-1 cadherins and applying a method statistically more robust than NJ, such as ML. In contrast to previous works that only use EC1-repeat to analyze cadherins evolution, using the full-length EC in the present work we could reclassify CDH13 into type I cadherins, suggesting their reliable evolution.

Furthermore, the present analysis suggests that CDH13 had appeared before Tetrapoda during evolution course, such as it was expected for their hypothesized regulatory role. Also, DSCs seem not to be mammalian exclusive desmosomal cadherins, as previously though.

We found that EC and CDs follow different evolutionary models, which supports the hypothesis that they have separately evolved.

Finally, we confirmed the homology relations of many putative uncharacterized proteins with the well characterized and reviewed sequences.

Undoubtedly, further genomes sequencing together with more cadherin protein structures elucidations will help to clarify the emergence, evolution and specialization of members of this superfamily.

5. Acknowledgements
This work was supported by a grant from Universidad Nacional de Entre Ríos (UNER), PID 6088-1. We thank Marisa Farber, Gustavo Parisi and Ariel Amadio for their helpful suggestions and comments. Also, we are very grateful to the Laboratory of Signals and non-Linear Dynamics from UNER, by allowing us to use their cluster of computers.

6. References