PROTEINS: Structure, Function, and Bioinformatics 65:808–815 (2006)

PREDICTION REPORT

Identification of New Claudin Family Members by a Novel PSI-BLAST Based Approach With Enhanced Specificity

Jun Wu,[1] Gerd Helftenbein,[1]* Michael Koslowski,[2] Ugur Sahin,[2] and Özlem Tureci[1]
[1]*Ganymed-Pharmaceuticlas AG, Freiligrathstraße 12, 55131 Mainz, Germany*
[2]*Department of Internal Medicine, Division for Experimental and Translational Oncology, Johannes Gutenberg University, Obere Zahlbacherstr.63, 55131 Mainz, Germany*

ABSTRACT **In an attempt to develop a novel strategy for the identification of new members of protein families by *in silico* approaches, we have developed a semi-automated procedure of consecutive PSI-BLAST (Position-Specific-Iterated Basic Local Alignment Search Tool) searches incorporating identificiation as well as subsequent validation of putative candidates. For a proof of concept study we chose the search for novel members of the claudin family. The initial step was an iterated PSI-BLAST search starting with the PMP22_Claudin domain of each known member of the claudin family against the human part of the RefSeq Database. Putative new claudin domains derived from the converged list were evaluated by a validating PSI-BLAST in which each sequence was assessed for finding back the starting set of known claudin domains. The local PSI-BLAST searches and validation were automated by a set of PERL scripts. With this strategy a total of three additional putative claudin domains in three different proteins were identified. One of them was subjected to further characterization and was shown to exhibit claudin-like features in terms of protein structure and expression pattern. The strategy we present is an efficient and versatile tool to identify novel members of domain-sharing protein families. Low rates of false positives achieved by inclusion of a validation step into the *in silico* procedure make this strategy particularly attractive to select candidates for subsequent labor-intensive wet bench characterization. Proteins 2006;65:808–815.** © 2006 Wiley-Liss, Inc.

Key words: PMP22_Claudin; PERL; domain; iterative

INTRODUCTION

One major objective of *in silico* approaches in molecular biology of the post-genomic era is the functional characterization of novel proteins, especially those predicted from genome annotation projects. To this aim, well-characterized proteins are subdivided in functional structures also known as domains, which are able to work as independent units within the protein. Because domains derived from different proteins of the same functional family usually also share a high degree of sequence similarity, sequence comparisons are well accepted in *in silico* approaches for the identification of new domain family members. The most popular sequence comparison tool based on local sequence alignments is the BLAST algorithm.[1] This algorithm aligns a single query sequence within an acceptable time span with each entry of a sequence database comprising up to millions of sequences.

The sensitivity of the basic BLAST search tool is limited by the fact that only a single sequence can serve as a query and only those sequences that share sufficient homology to that particular sequence will result in a BLAST hit. In biological reality, however, most sequences characteristic for a given domain are not well represented by the sequence of only one family member but by a virtual consensus sequence, which is derived from all sequences within that family. Although consensus sequences improve the sensitivity of BLAST searches when used as a query, they still have the limitation that for any given position only one defined amino acid is permitted. This restriction is circumvented and sensitivity is enhanced by modern sequence comparison tools, which use sequence models or sequence profiles such as HMM (Hidden-Markov-Model)[2,3] rather than single sequences for the representation of domain sequences. The most comprehensive sets of domain sequence models in combination with sequence alignment tools for the characterization of novel sequences are found in SMART[4] or Pfam.[5]

Jun Wu and Gerd Helftenbein contributed equally to this work.

Jun Wu's current address is: Shanghai Center for Bioinformation Technology, Qinzhou Rd, 100 Shanghai, China.

*Correspondence to: Gerd Helftenbein, Ganymed Pharmaceuticals AG, Freiligrathstrasse 12, D-55131 Mainz, Germany.
E-mail: g.helftenbein@ganymed-pharmaceuticals.com

Received 1 June 2006; Revised 18 July 2006; Accepted 2 August 2006

Published online 4 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21218

Profiles of sequences not only serve as searchable databases but can also be used themselves as query against natural sequences. One such tool is the PSI-BLAST (Position-Specific-Iterated Basic Local Alignment Search Tool) algorithm developed by Altschul et al. in 1997[6] and improved by Schäffer et al. in 2001.[7] It is based on the BLAST algorithm for the alignment of an initial query sequence against a sequence database and uses this first BLAST output for the calculation of a sequence profile that can be repeatedly subjected to the BLAST search. The main advantage of PSI-BLAST compared to the classical BLAST tool is that such modeling by iterative cycles of improvement increases sensitivity. As a downside, such sequential alterations bear the risk of alienation of the sequence profile. Accordingly, critical validation of final PSI-BLAST outputs for their relatedness to the initial query sequence in terms of homology and maybe functionality is required.

Proteins of the claudin family are integrative components of tight junction complexes in epithelia and endothelia of different tissues.[8] Tight junctions are complex paracellular barriers also known to be involved in signaling cascades controlling cell growth and differentiation.[9] The multigene family of claudins comprises more than 20 members that typically encode 20–27 kDa proteins. All claudins are tetraspanins and share the feature of four transmembrane domains with two short extracellular loops and an intracellular carboxyterminus haboring a PDZ binding motif.[10] The four transmembrane domains as a whole are a basic structural component of a PMP22_Claudin domain which is listed in PFAM (accession number PF00822) as a common motif for claudins, PMP-22 and EMP-1, -2, and -3.[11–13] The modular character of that domain makes it a perfect proof-of-concept model for the identification of new claudin domains with the number of transmembrane domains as an appropriate indicator for the quality of a putative PMP22_Claudin domain.

Here we report a novel strategy to identify putative claudin domains within a comprehensive database of human proteins by a workflow of subsequent PSI-BLAST searches automated by a set of PERL scripts. The novelty of this concept lies in the intrinsic evaluation of the PSI-BLAST outputs by the capability of each BLAST hit to find back the initial claudin domains in a second validating PSI-BLAST.

The potential of this strategy is underlined by the discovery of three novel proteins with a putative PMP22_Claudin domain as a central component.

METHODS

Workflow Description

The work described here is based on the human part of the RefSeq database, Release 8[14] and the standalone BLAST package BLAST-2.2.6 downloaded from the NCBI ftp-Server at ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.6/and installed locally on a Linux machine.

The starting set of human claudin proteins was downloaded from NCBI by an Entrez search with the command "claudin [Protein name] AND "Homo sapiens" [Organism] AND srcdb_refseq [PROP]". This led to the list of 28 claudins, which are shown in Table I and served as an entry point for the entire workflow. The workflow consists of a set of 15 scripts organized in 3 modules and was written in PERL for automation. The set of scripts is available from the authors upon request.

TABLE I. Initial Claudin List[a]

NP_001296	Claudin 4
NP_001001346	Claudin 20
NP_955360	Claudin 8
NP_919260	Claudin 23
NP_878268	Claudin 10 isoform a
NP_001298	Claudin 7
NP_683763	Claudin 19
NP_652763	Claudin 14
NP_612438	Claudin 15
NP_078876	Claudin-like protein 24
NP_067018	Claudin 6
NP_066192	Claudin 9
NP_005593	Claudin 11
NP_066924	Claudin 1
NP_065117	Claudin 2
NP_057453	Claudin 18 isoform 1
NP_055158	Claudin 15
NP_036263	Claudin 17
NP_036262	Claudin 14
NP_036261	Claudin 12
NP_008915	Claudin 10 isoform b
NP_006571	Claudin 16
NP_003268	Claudin 5
NP_001297	Claudin 3
NP_149098	Claudin 12
NP_001013765	Claudin domain-containing protein
NP_001002026	Claudin 18 isoform 2
XP_210581	Claudin 22

[a]List of human claudins and RefSeq accession numbers that were used as the initial set for the validating PSI-BLAST workflow.

Module 1 (sequence selection) comprises 4 scripts that create Genpept and FASTA files of the Entrez output. The FASTA sequences are then concatenated to a multifasta file. The multifasta file is analyzed by SMART and the last script extracts the PMP22_Claudin domain sequences to single FASTA files. The PMP22_Claudin domains of all human claudins, which are created by this module, are subjected to module 2.

Module 2 (exploring PSI BLAST) is the most complex one and consists of 8 scripts. The first one uses each PMP22_Claudin domain sequence from module 1 for a PSI-BLAST search against the human RefSeq mRNA database. The output is parsed, collected in one file, and made non-redundant. After removing the 28 prototypic input sequences, the remaining putative PMP22_Claudin domains are collected and each sequence is written to a FASTA file for the validating PSI-BLAST.

Module 3 (validating PSI-BLAST) is represented by the last 3 scripts, which perform PSI-BLAST searches of each putative domain against human RefSeq mRNAs with subsequent evaluation of the output. As a criterion for high quality hits, the number of recovered initial

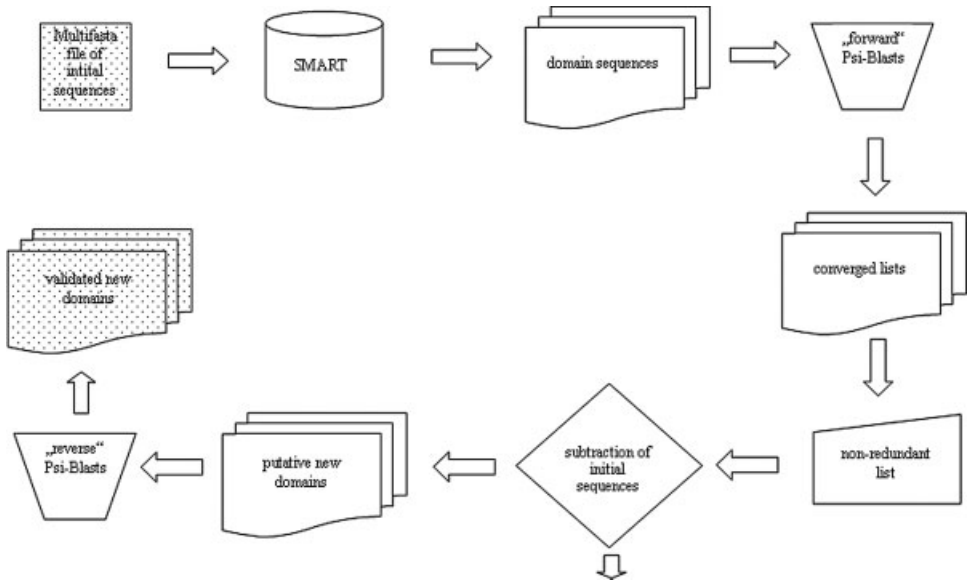

Fig. 1. Schematic representation of the validating PSI-BLAST workflow for the identification of new protein family members. Start and endpoint are marked by dots.

sequences is assessed and each putative domain sequence is extracted.

RT-PCR

Extraction of RNA and preparation of oligo-dT primed first strand cDNA derived from human tissues and cell lines was done as described previously.[15] For a typical reaction 0.5 μl of first-strand cDNA was amplified with transcript-specific oligonucleotides (10 pmol each) using 1 unit of HotStarTaq polymerase (Qiagen, Hilden, Germany) in a 30-μl reaction. Gene-specific primers were designed for the putative new claudins (XP_211287 – sense GTG CAG CAC CTC ATC TTG CTG starting at position 880 and antisense CAG GTG CGA GTA GCT GAT GTA starting at position 1083; XP_208930 – sense ATT CTG CTG CCG CTC AGC CTG starting at position 20 and antisense ACG CTA TGT AGA CGC TGA TCC starting at position 176; RefSeq: NP_689603 – sense ATG AGT GGC ACC TGC TTC ATT from position 522 to antisense CGG CAT AAG TAC AGA GGG AAA at position 713 of the published mRNA).

Amplification was performed using a TRIO thermoblock (Biometra, Göttingen, Germany). After a 15 min activation of HotStarTaq polymerase at 95°C for hot start induction, 35 cycles of PCR were performed with 30 sec at the annealing temperature of 66°C, 30 sec at 72°C, and 30 sec at 94°C, with a final elongation step at 72°C for 6 min. A 20 μl-aliquot of each reaction was size-fractionated on a 1.5% agarose gel, visualized by ethidium bromide staining, and assessed for the amplification of a product of expected size.

Sequence Alignment

The ClustalW alignment program[16] available at the EBI webserver at http://www.ebi.ac.uk/clustalw/ was used for pairwise as well as multiple alignment and phylogenetic tree calculation.

RESULTS

Strategy

The workflow process we developed has the objective to discover new candidates of a defined protein family within a comprehensive protein database. To this aim, all human proteins were downloaded from the Refseq database (Release 8)[14] by Entrez searches at the NCBI Webserver in Genpept and FASTA format. For performance reasons the BLAST package BLAST-2.2.6 was also downloaded from NCBI ftp-server and installed as standalone tool on a standard Linux machine.

The complete workflow is schematically shown in Figure 1. The initial step was an Entrez search in the RefSeq database for mRNAs of the protein family we wanted to expand, namely human claudins. The resulting 28 hits (Table I) were downloaded in GenPept format and the corresponding FASTA sequences were collected in a Multifasta file by a first PERL script. The Multifasta file was analyzed by the SMART tool[4] for presence and location of each PMP22_Claudin domain. A second PERL script conducted automated parsing of the PMP22_Claudin domain sequence of each protein into a new FASTA file that served as the entry point of the following PSI-BLAST search. Out of 28 protein sequences 20 were predicted by SMART to encompass a PMP22_Claudin domain. Each of the 20 sequences of this prototypic reference set was subsequently used for PSI-BLAST searches against the human protein part of the RefSeq database in a script driven mode allowing selection of different E-value thresholds. All runs were conducted until the search converged, meaning that no new hits were found in the last

TABLE II. PSI-BLAST Results[a]

Exploring E-value	Non-redundant His	Validating E-value		
		0.005	0.05	0.5
0.005	17	NP_689603 (22) XP_211287 (24)	NP_689603 (24) XP_211287 (24)	NP_689603 (24) XP_211287 (24) NP_653233 (24) NP_722545 (24) NP_872354 (24) NP_653276 (21)
0.05	23	NP_689603 (22) XP_211287 (24) XP_208930 (21)	NP_689603 (24) XP_211287 (24) XP_208930 (21)	NP_689603 (24) XP_211287 (24) NP_653233 (24) NP_722545 (24) NP_872354 (24) XP_208930 (22) NP_653276 (21)
0.5	75	NP_689603 (22) XP_211287 (24) XP_208930 (21)	NP_689603 (24) XP_211287 (24) XP_208930 (21)	NP_689603 (24) XP_211287 (24) NP_653233 (24) NP_722545 (24) NP_872354 (24) XP_208930 (22) NP_653276 (21)

[a]Results of the validating PSI-BLAST workflow dependent on the E-value threshold used for the exploring and validating PSI-BLAST. Validated Hits are represented by RefSeq accession numbers followed by the number of initial claudins found in the validating PSI-BLAST.

iteration. The resulting hit lists were automatically combined and checked for redundancy. After subtraction of the original 28 claudins the remaining list included different numbers of hits depending on the E-value threshold used for the PSI-BLAST profile calculation. With an E-value threshold of 0.005 a total of 17 hits were found, whereas a threshold value of 0.5 resulted in 75 hits.

Next, a different set of scripts created the corresponding FASTA file of each hit list that was used for a validating PSI-BLAST, again with different E-value thresholds. We employed this validating PSI-BLAST for each putative PMP22_Claudin by automated assessment of the converged list of each search for the presence of the 28 claudins serving as reference. The number of prototypic claudins recovered at this step should correlate in principle with the proximity of the putative domain to the known domains in terms of sequence similarity. Depending on the different E-value thresholds chosen for exploring and validating PSI-BLAST, this resulted in further collapsing of the list to 6 or 7 hits as shown in Table II.

Three main conclusions can be drawn from this data. First, although changing the exploration E-value threshold from 0.05 to 0.5 does increase the number of initial hits dramatically, this does not substantially affect the output of the validating PSI-BLAST. This can be explained by the relative unspecificity of the exploring search, which leads to a large number of unrelated false positives. Second, an increase of the exploration E-value from 0.005 to 0.05 yields XP_208930 as only additional hit, which is approved by finding back initial claudins at a high ratio even at the most stringent validating search conditions (E-value of 0.005). This suggests that an exploration E-value threshold of 0.05 is appropriate to yield the most comprehensive hit list with a high chance of passing the most stringent validating PSI-BLAST conditions. Third, those proteins which eventually pass the filter of the validating PSI-BAST have high rates (at least 21 out of 28) of finding back to the prototypic claudins we used as reference for validation. Apparently, the sequence profile obtained by the refinement process renders a clear-cut distinction. Either the profile "runs into" the claudin profile, resulting in a validated hit list of nearly all claudins or it not even touches the claudins.

Taken together this data suggests that an exploration E-value of 0.05 is sufficient for a comprehensive list of candidates and that confirmation should be performed with high stringency using at a validation E-value threshold of 0.005. Validation E-values up to 0.5 appear to be too unspecific and additional hits they yield are less likely related to claudins.

Analysis of Novel Putative Claudin Family Members

Applying the E-value combination of 0.05/0.005 we determined as optimal, we identified three hits NP_689063, XP_211287, and XP_208930 as putative novel members of the claudin family.

All three were predicted proteins of the RefSeq project with no further experimental data available. At the time they were picked up by this workflow, the biological functions of all three proteins were completely unknown and potential links to the claudin-family had not been reported previously.

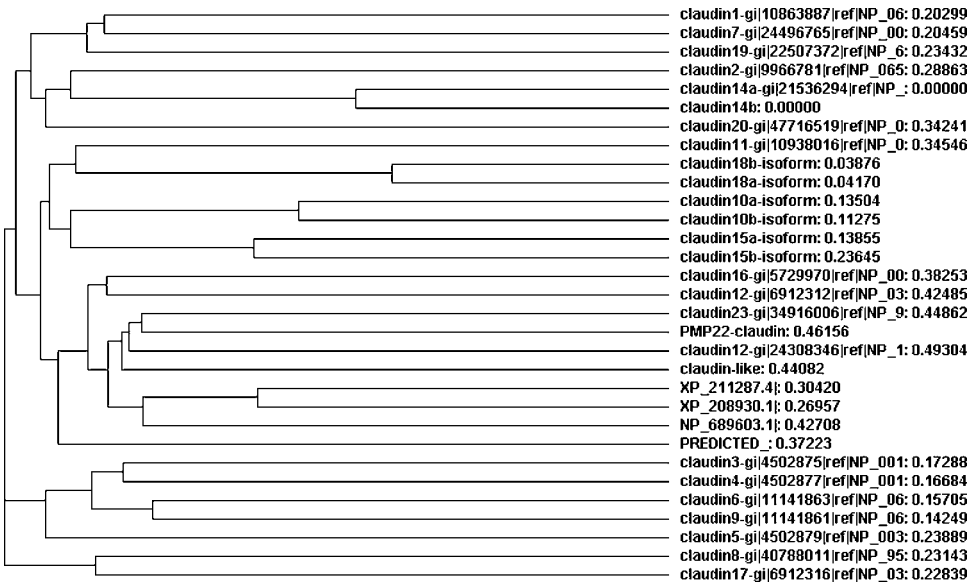

Fig. 2. Phylogenetic tree including distances of all claudin sequences used as entry and reference sets for the PSI-BLAST searches together with the putative new claudins named XP_211287, XP_208930, and NP_689603. Calculation was done by the ClustalW alignment program at EBI (http://www.ebi.ac.uk/clustalw/). The putative new sequences cluster together with highest homology to claudin12, claudin23, claudin-like protein 24 NP_078876 and PMP22-claudin domain-containing protein NP_001013765.

Analysis of XP_211287 as well as NP_689603 with TMPred[17] disclosed four transmembrane domains as typical substructure of PMP22_Claudin domains and supported that these new sequences represent claudin family members. XP_208930 appeared to be fragmentary because the predicted open reading frame encoded a protein of 122 amino acids encompassing only 3 transmembrane domains and thus too short to represent a PMP22_Claudin domain of about 180 amino acids.

We performed sequence similarity analysis of all 28 claudin domains, including the 3 putative new ones using ClustalW,[16] and visualized the output in a phylogenetic tree view. As shown in Figure 2 and as expected, the putative claudin family members XP_211287, XP_208930, and NP_689603 discovered by us blended perfectly into the claudin domain family tree. They clustered together as one branch with highest relatedness to claudin12 NP_036261, claudin23 NP_919260, claudin-like protein 24 (CLP24; NP_078876), and PMP22-claudin domain-containing protein NP_001013765. Claudin23 is reported to be expressed in gastric cancer[18,19] and expression of CLP24 has been described in lung, heart, kidney, and placental tissues.[20]

Characterization of XP_211287

We chose XP_211287, which has a HSP (High-scoring Segment Pair) between amino acid 6 and 196, for detailed analysis. As compared to analogous stretches in the two other sequences we discovered, this stretch of 190 amino acids found back to the highest number of reference claudins (24 out of 28), suggesting that it has the closest relatedness to known claudins.

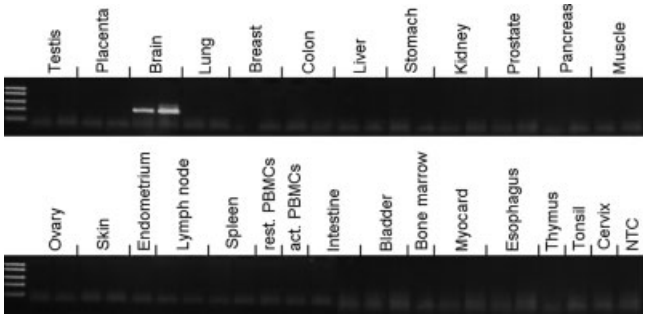

Fig. 3. End-point RT-PCR analysis of XP_211287 expression. Experimental conditions are described in the Methods section. The majority of tissues are represented by two independent specimens designated as 1 and 2. Only two cDNA samples derived from human brain yielded the expected 203 bp amplification product. PBMCs, peripheral blood mononuclear cells; rest., resting; act., activated; NTC, non template control.

To investigate expression of the gene encoding XP_211287 and study its distribution in human tissues, we performed RT-PCR analysis of a comprehensive set of primary human tissue specimen derived from different human organs. We found that expression of this gene is highly selective for and restricted to human brain tissue, whereas it is not detected in any other tissue type we investigated (see Fig. 3).

We assumed that commitment of members of the same protein family to a defined cell lineage might be a better hint for related function as f.e. a higher extent of mere protein sequence similarity. This triggered a literature survey for claudin family members already known as expressed in or restricted to brain tissue. We found references for brain specific expression of claudin1[21,22] and of claudin5,[21,23] which is described as a component of tight junctions of the blood brain barrier.

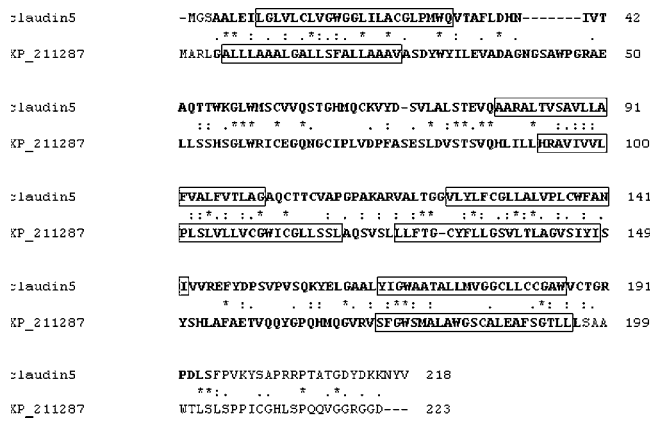

Fig. 4. Sequence alignment of claudin5 cDNA NP_003268.1 and XP_211287.4 was done by the ClustalW alignment program provided at the EBI webserver. Identical amino acids are indicated by a star, conserved substitutions are indicated by two dots and semi-conserved substitutions by one dot. The putative claudin domain of XP_211287 (6 to 196) and the known domain of claudin5 (4 to 195) are marked bold. Transmembrane domains are boxed in both cases. Structural similarity of both proteins is obvious.

Direct sequence alignment of claudin5 and XP_211287 showed that with regard to protein sequence, the gene products exhibit a moderate similarity mainly in the region of the predicted claudin domain (marked bold), which comprises about 85% of the entire protein (see Fig. 4). As XP_211287 has no PDZ-binding motif, the short C-terminus of both proteins is less similar. A comparison in terms of structure, however, revealed striking similarities. First, both gene products have a comparable length. Second, the patterns in which the four transmembrane domains are arranged in these proteins resemble each other (see Fig. 4). In both cases there is one transmembrane domain located at the N-terminus followed by a stretch of about 50 amino acids predicted by TMHMM to be located at the outer cell surface followed by a set of three transmembrane domains in close proximity to each other. Third, no additional domains were predicted by SMART or Pfam.

These observations further support that XP_211287 is a novel claudin family member and most likely functionally similar to claudin5.

DISCUSSION

Functional classification of proteins is one of the major objectives in molecular biology not only for the characterization of known proteins but also for prediction of a biological function for hypothetical ones derived from gene prediction programs. The RPS-BLAST,[24] for example, allocates known protein domain models to a given sequence by using the BLAST algorithm. Moreover, tools for functional classification are valuable for the explorative systematic search for novel members of a defined protein family.

Accordingly, improved *in silico* approaches for identification and expansion of functional domains within proteins are still in the focus of interest.

A major disadvantage of basic alignment tools such as classical BLAST is their restriction to one sequence as a query. Even if a consensus sequence derived from multiple sequence alignments is used, it still features one defined amino acid at each position. In contrast, a model such as HMM[2,3] takes the relative frequency of every amino acid at each position into account and thus provides a better representation of an alignment. The inclusion of such sequence models as query sequences in the classical BLAST tool led to the development of PSI-BLAST in 1997.[6] Even with these developments, however, a general problem of *in silico* approaches, namely, the high rate of false positives remains.

The approach to increase specificity we have chosen is conceptually new, as it inaugurates the evaluation of PSI-BLAST results by PSI-BLAST itself. To this aim, an exploring PSI-BLAST search to identify putative new homologues is combined with a subsequent validating PSI-BLAST as a first evaluation of the preliminary results. To be accepted as a hit, a predicted sequence has to find back those sequences in a validating PSI-BLAST search, which were used as the entry point and reference set for the first exploring PSI-BLAST search.

Altered PSI-BLAST strategies have been developed to further increase sensitivity.[25,26] In particular for protein annotation comprehensive tool boxes such as MyHits[27] or DOMAINATION,[28] which use protein motif scans and motif based searches, are available.

However, to our knowledge no attempt is described to validate the output of such a search in terms of specificity within one comprehensive workflow and in neither of them a validating PSI-BLAST step is conducted.

In cases analogous to our proof of concept model claudin, where each member of a given family is used in an independent PSI-BLAST search and the combined and non-redundant result of each search is evaluated by a validating search, the procedure may become rather complex. Therefore, we automated the entire workflow with a versatile set of PERL scripts. Variable parameters in these scripts are the number of sequences from which each single exploring PSI-BLAST starts and the E-value thresholds both for exploring and validating searches.

The set of sequences used initially as entry for the exploring BLAST and later as reference for the validating BLAST can be determined either by Entrez queries in the RefSeq database at NCBI using protein family names as keywords for searching or by the alignment based SMART function of listing all proteins comprising a domain of interest. In our proof of concept example we combined both approaches.

An Entrez search for human claudins identified 28 sequences. However, only in 20 of these a PMP22_Claudin domain was predicted by SMART, suggesting that the SMART sequence model may not represent all claudins. We used the 20 SMART-defined sequences as entry set and the 28 claudins found by ENTREZ as reference set for final validation.

We identified three proteins with putative PMP22_claudin domains whereby assigning them to the claudin family. At the time we picked them up, the biological functions of all three proteins were completely unknown and potential

links to the claudin-family had not been reported previously. Their assignment to the claudin family suggests that these proteins share functional features with claudins, which are typically involved in the formation of tight junctions. The key function of tight junctions is maintenance of a luminal barrier, paracellular transport, and signal transduction. Disruption of tight junctions can cause loss of cell polarity, resulting in an abnormal influx of fluids and soluble factors.

In the meantime, a PMP22_Claudin domain was assigned to XP_211287 by SMART. Moreover, NP_689603 is listed in UniProt[29] and represented by two TrEMBL entries (Q8N6N4, Q8NBL3), which are annotated as claudins although SMART is not able to detect a PMP22_Claudin domain even with activated outlier search conditions. This not only validates our approach retrospectively, but also shows its value for the identification of yet uncharacterized hypothetical proteins not included in annotated databases.

Further analysis of XP_211287, which has the highest rate of finding back other claudins, disclosed that expression of this molecule is highly restricted to human brain tissue.

From all known claudins only claudin5 and claudin1 are described to be expressed in human brain.[23] Downregulation of both claudins together with occludin are key molecular abnormalities responsible for the increased permeability of tumor endothelial tight junctions in brain tumors and consecutive edema of brain tissue.[21]

Surprisingly, knockout of claudin5 was reported to result in a selective increase in paracellular permeability of small molecules rather than in the expected general breakdown of tight junctions.[30] The authors concluded from this finding that there is another yet unidentified claudin or claudin-homolog involved in the architecture of tight junctions of the blood-brain barrier. The tissue specific expression of XP_211287 together with a putative role as claudin makes this hypothetical protein a promising candidate for that missing claudin. Further analysis including immunolocalisation of tagged XP_211287 in transfected cells, immunohistochemical staining of brain tissue sections with antibodies directed against endogenous XP_211287, and knock-out systems may assist to confirm this hypothesis. In summary, we have shown the identification of yet unidentified additional members of the claudin domain family by a semi-automated *in silico* approach. The validating PSI-BLAST workflow described here is capable of not only finding related members of protein families with high sensitivity, but also validating them by application of a "finding back to the initial sequences" criterion and thus providing specificity as well. As shown by additional yet unpublished results, this approach can be extended in principle to other protein families. Instead of the Refseq database, which proved useful and appropriate for the claudin project, other applications may require the access of other databases or combinations of databases for a higher sensitivity.

The approach, we presented here is particularly suited for those protein families that are characterized by the presence of one well-defined domain.

REFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
2. Churchill GA. Stochastic models for heterogeneous DNA sequences. Bull Math Biol 1989;51:79–94.
3. Eddy SR. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol 1995;3:114–120.
4. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA 1998;95:5857–5864.
5. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32:D138–D141.
6. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
7. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
8. Furuse M, Fujita K, Hiiragi T, Fujimoto K, Tsukita S. Claudin-1 and -2: novel integral membrane proteins localizing at tight junctions with no sequence similarity to occludin. J Cell Biol 1998; 141:1539–1550.
9. Gonzalez-Mariscal L, Betanzos A, Nava P, Jaramillo BE. Tight junction proteins. Prog Biophys Mol Biol 2003;81:1–44.
10. Morita K, Furuse M, Fujimoto K, Tsukita S. Claudin multigene family encoding four-transmembrane domain protein components of tight junction strands. Proc Natl Acad Sci USA 1999; 96:511–516.
11. Taylor V, Welcher AA, Program AE, Suter U. Epithelial membrane protein-1, peripheral myelin protein 22, and lens membrane protein 20 define a novel gene family. J Biol Chem 1995;270: 28824–28833.
12. Marvin KW, Fujimoto W, Jetten AM. Identification and characterization of a novel squamous cell-associated gene related to PMP22. J Biol Chem 1995;270:28910–28916.
13. Ben Porath I, Benvenisty N. Characterization of a tumor-associated gene, a member of a novel family of genes encoding membrane glycoproteins. Gene 1996;183:69–75.
14. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet 2000;16:44–47.
15. Grunwald C, Koslowski M, Arsiray T, Dhaene K, Praet M, Victor A, Morresi-Hauf A, Lindner M, Passlick B, Lehr HA, Schafer SC, Seitz G, Huber C, Sahin U, Tureci O. Expression of multiple epigenetically regulated cancer/germline genes in nonsmall cell lung cancer. Int J Cancer 2006;118:2522–2528.
16. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Methods Enzymol 1996;266:383–402.
17. Hofmann K, Stoffel W. TMbase—a database of membrane spanning proteins segments. Biol Chem Hoppe Seyler 1993;374:166–170.
18. Gress TM, Muller-Pillasch F, Geng M, Zimmerhackl F, Zehetner G, Friess H, Buchler M, Adler G, Lehrach H. A pancreatic cancer-specific expression profile. Oncogene 1996;13:1819–1830.
19. Katoh M, Katoh M. CLDN23 gene, frequently down-regulated in intestinal-type gastric cancer, is a novel member of CLAUDIN gene family. Int J Mol Med 2003;11:683–689.
20. Kearsey J, Petit S, De Oliveira C, Schweighoffer F. A novel four transmembrane spanning protein, CLP24. A hypoxically regulated cell junction protein. Eur J Biochem 2004;271:2584–2592.
21. Papadopoulos MC, Saadoun S, Binder DK, Manley GT, Krishna S, Verkman AS. Molecular mechanisms of brain tumor edema. Neuroscience 2004;129:1011–1020.
22. Swisshelm K, Macek R, Kubbies M. Role of claudins in tumorigenesis. Adv Drug Deliv Rev 2005;57:919–928.
23. Virgintino D, Errede M, Robertson D, Capobianco C, Girolamo F, Vimercati A, Bertossi M, Roncali L. Immunolocalization of tight junction proteins in the adult and developing human brain. Histochem Cell Biol 2004;122:51–59.

24. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 2005;33:D192–D196.
25. Larson SM, Garg A, Desjarlais JR, Pande VS. Increased detection of structural templates using alignments of designed sequences. Proteins 2003;51:390–396.
26. Anand B, Gowri VS, Srinivasan N. Use of multiple profiles corresponding to a sequence alignment enables effective detection of remote homologues. Bioinformatics 2005;21:2821–2826.
27. Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Falquet L. MyHits: a new interactive resource for protein annotation and domain identification. Nucleic Acids Res 2004;32: W332–W335.
28. George RA, Heringa J. Protein domain identification and improved sequence similarity searching using PSI-BLAST. Proteins 2002;48:672–681.
29. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2004;32: D115–D119.
30. Matter K, Balda MS. Holey barrier: claudins and the regulation of brain endothelial permeability. J Cell Biol 2003;161:459, 460.