Iranian Journal of Veterinary Research, Shiraz University, Vol. 12, No. 3, Ser. No. 36, 2011

Comparing various attributes of prolactin hormones in different species: application of bioinformatics tools

Ebrahimi, M.

Department of Biology, School of Basic Sciences, University of Qom, Qom, Iran and Bioinformatics Research Group, Green Research Center, University of Qom, Qom, Iran

Correspondence: M. Ebrahimi, Department of Biology, School of Basic Sciences, University of Qom, Qom, Iran and Bioinformatics Research Group, Green Research Center, University of Qom, Qom, Iran. E-mail: Ebrahimi@qom.ac.ir

(Received 13 Jun 2010; revised version 22 Jan 2011; accepted 5 Feb 2011)

Summary

Prolactin is mainly secreted by the anterior pituitary and is able to stimulate mammary gland development and lactation in mammalians. Although prolactins share a common ancestral gene encoding, they show species specific characteristics and their efficiency may be different in various mammals. The importance of protein structures of all sequences of this hormone have been studied by various bioinformatics algorithms. The results showed bioinformatics tools and modeling methods can be used to identify the species specificity of prolactin hormones in animals with an acceptable precision rate. Based on the author's knowledge, this is the first report on the structural variation of prolactin hormones by specific structural protein features. Gain ratio model acquired the best accuracy and performance among the algorithms applied here and can be used on similar proteins. The counts and the frequencies of dipeptides were the most important protein attributes in this regard. It has also been reported here that feature selection or attribute weighting can be used to select the most important protein attributes and to reduce the burden of processing equipment. The new findings presented here open up new windows in understanding the characteristics of prolactin hormones and also pave the way to engineer more efficient hormones by using various mutagenesis tools such as site directed mutagenesis.

Key words: Prolactin, Modeling, Bioinformatics, Data mining

Introduction

Prolactin is mainly secreted by lactotrophic cells of the anterior pituitary and its secretion is mainly controlled by inhibitory factors originating from the hypothalamus (Pariante, 2008). The prolactin gene is regulated at the transcriptional level by two distinct promoters. The proximal promoter, also referred to as the pituitary promoter, covers ~5 kb upstream of the transcription site, in which the 250 bp just before the Cap (capping of polymerase on RNA) site (in exon 1 b) are necessary and sufficient for transcription (Hiyama *et al*., 2009). The second promoter, referred to as the extra-pituitary promoter, includes ~3 kb upstream of exon 1 a (itself located ~5.8 kb upstream for the initiation site) (Swaminathan *et al*., 2008). Depending on promoter usage, prolactin mRNAs differ in length by 134 bp, but they encode identical mature protein. Posttranslational modifications are not required for the hormone to be fully active. In fact, posttranslational modifications are more often detrimental than beneficial to prolactin bioactivity (Nichols and Green-Church, 2009).

Prolactin shares high structural and functional similarity with two other polypeptide hormones, growth hormone and placental lactogen (Ben-Jonathan *et al*., 2008; LaPensee *et al*., 2009). It is thought that the genes encoding these proteins evolved from a common ancestral gene by duplication. More recently, newly identified proteins such as proliferin, proliferin-related-protein, somatolactine, or several prolactin-like proteins have been added to this family based on sequence similarities (Liu *et al*., 2009). More than 300 separate biological activities have been attributed to prolactin, and can be subdivided into the

following categories: functions linked to reproduction, endocrinology and metabolism, control of water and electrolyte balance, growth and development, brain and behavior, and finally, immunoregulation and protection (Trott *et al*., 2008).

Data mining problems often involve hundreds or even thousands of variables (Ye *et al*., 2009). Fitting a model such as a decision tree or item set mining to a set of variables this large may require more time than is practical (Gromiha and Yabuki, 2008). Usually, many attributes determine the different characteristics of a protein molecule. As a result, the majority of time and effort spent in the model-building process involves determining which variables to include in the model. Various models such as attribute weighting (or feature selection) allow the variable set to be reduced in size, creating a more manageable set of attributes for modeling (Zhu *et al*., 2010). The decision tree algorithm predicts the value of a discrete dependent variable with a finite set from the values of a set of independent variables (Dancey *et al*., 2007). A decision tree is constructed by looking for regularities in data, determining the features to add at the next level of the tree using an entropy calculation, and then choosing the feature that minimizes the entropy impurity (Gromiha, 2007). Several well-known decision tree algorithms are available (Huang *et al*., 2009). To better understand the features that contribute to structural differences between prolactin hormones in various species, it is necessary to identify the main features responsible for this valuable characteristic. Herein we have used various clustering, screening, item set mining and decision tree models to determine the most important features responsible for prolactin hormones in various species.

Materials and Methods

All available protein sequences of prolactin hormone (112 so far) from various animals (alligator, bovine, camel, carp, cat, catfish, chicken, deer, eel, elephant, flounder, goat, goldfish, sheep, possum, human, pig, mink, tilapia, salmon, trout, goat, mouse, rat, flounder, pigeon, sea bass, sea bream, turkey, whale, turtle, lungfish, hamster, toad and monkey) have been extracted from the UniProt knowledgebase (Swiss-Prot and TrEMBL; www.expasy.org) database. Eight hundred and ninety nine protein features such as length, weight, isoelectric point, count and frequency of each element (carbon, nitrogen, sulfur, oxygen and hydrogen), count and frequency of each amino acid, count and frequency of negatively charged, positively charged, hydrophilic and hydrophobic residues, count and frequency of dipeptides, number of α-helix and β-strand and other secondary protein features, as well as bond angle, bond length, dihedral angle and other tertiary protein features were extracted. All features were classified as continuous variables, except for the N-terminal amino acid and the type of organisms which were classified as categorical. A dataset of these protein features was imported into Clementine software (Clementine_NLV-11.1.0.95; Integral Solution, Ltd.), null data for the type of organism were discarded and set as the output variable and the other variables were set as input variables. The same database was imported into RapidMiner software (RapidMiner 5.0.001, Rapid-I GmbH, Stochumer Str. 475, 44227 Dortmund, Germany) and the type of organism set as the target or label attribute (when Item Set Mining model performed, no label or target attribute was set as this model requires so).

To identify the most important features contributing to the type of prolactin hormone in different animals, various screening models (anomaly detection model, feature selection algorithm or attribute weighting), clustering models (K-Means, TwoStep cluster), tree Induction models (with various criterion, C5.0, C5.0 with 10-fold cross validation and C&RT), Item Set Mining (FPGrowth) and Rule Induction model (10 fold cross-validation through stratified sampling) were employed as described previously (Ebrahimi *et al*., 2009). Whenever requested by the model, data were discretized by frequency; i.e. data were divided into 3 bins (ranges) with nearly equal frequencies in each class (low 0-0.3, mid 0.3-5 and high >0.5), and sometimes data were converted to nominal, and in some cases to binominal datasets.

Results

The statistical findings on prolactin hormones have been presented in Table 1. In 87.06% of proteins the N-terminal amino acid was Met; in 9.41, 1.18, 1.18, and 1.18% of proteins the same position was occupied by Pro, Lys, Leu, and Ile, respectively. Figure 1 is a web graph that illustrates the strength of the relationship between N-terminal amino acids and species. Met exhibits a strong relationship with most organisms (a thicker line shows a stronger relationship). Leu was the only N-terminal amino acid found in turtle, lungfish, camel, whale and alligator, while Ile, Lys and Pro were found solely in carp, elephant and toad.

Attribute weighting

As data should be normalized before running various weighting models (as it follows), it would be reasonable to expect that all weights be between 0 to 1. The results of attribute weighting algorithms have been presented in Appendix 2. The number of attributes with weight higher than 0.70 were: 14, 8, 15, 139, 26, 17, 7, 4, 332 and 3 in PCA, SVM, relief, uncertainty, gini index, Chi-squared, deviation, rule, gain ratio and info gain weighting models, respectively.

Item set mining

When FP-Growth was run on all attributes, more than 281 rules were created. The support of the rules went up to 99% for the nitrogen count when its value was high (higher than 0.5). When the weight of the proteins and the nitrogen count were high, the support lowered to 99%. When the half-life of mammals and hydrogen count were high, the supports were 89%. When the

Table 1: Rules induced by information gain criterion of rule induction model on numeric data

Rules	Organisms
If the freq of nitrogen ≤0.536 and count of negatively charged ≤0.523	Mouse
If the count of Arg (R) ≤0.602 and isoelectric point >0.169 and Ala-Trp ≤0.500	Rat
If the count of Asp-Gly >0.833	Salmon
If the count of Gln-Val >0.500	Alligator
If the count of Ser-Ser >0.611	Carp
If the count of Asp-Lys >0.500	Tilapia
If the count of Asp-Ile ≤0.250	Hamster
If Ala-Arg >0.833	Eel
If the count of Phe-Arg >0.500	Grey opossum
If the count of Cys-Leu >0.750	Seabream
If the count of Lys-Thr >0.833	Human
If the count Gly-Arg >0.833	Pig
If the count of hydrophobic residue >0.754	Sheep
If the count of Ala-Lys >0.833	Elephant
If the count of hydrogen >0.768	Horse
If the count of His-Ser >0.833	Chicken
If the count of Ala-Gln ≤0.167	Turkey
If the frequency of Asn-Gln >0.346	Camel
If count of oxygen >0.631	Mink
If the count of Ile-Gln >0.750	Whale
If the count of Cys-Asp >0.250	Turtle
If the count of Ala-Cys >0.500	Lungfish
If the count of Beta sheet ≤0.036	Toad
If the count of Glu-His >0.750	Cat
If the count of Ala-Met >0.750	Catfish
If the count of Glu-Val >0.250	Monkey
If the count of Ala-Leu >0.500	Goldfish
If the count of Ala-Asp >0.750	Rabbit
If the count of Ala-Asp >0.250	Deer
If the count of Beta sheet ≤0.429 then panda	Panda
Else flounder	

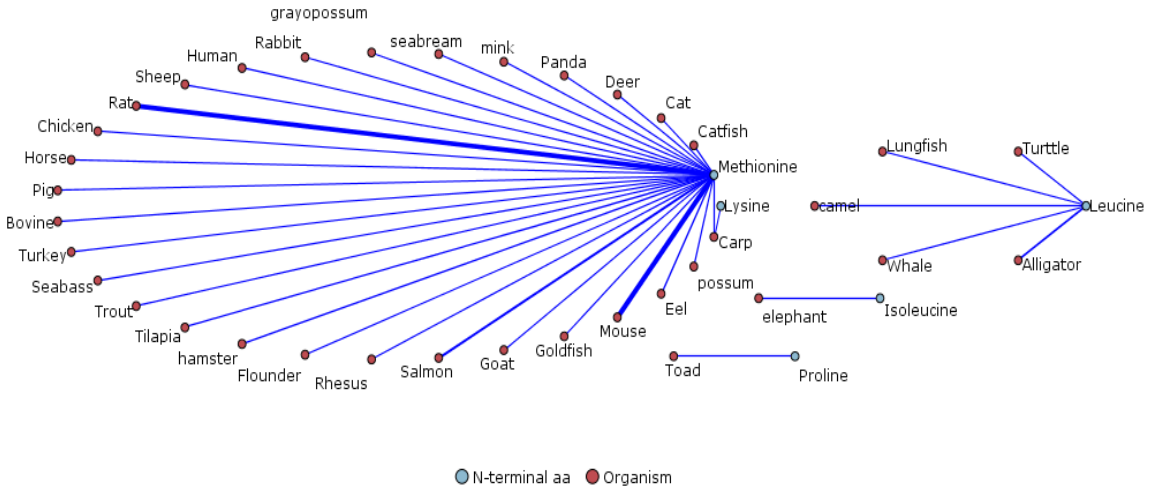

Fig. 1: Web graph of N-terminal amino acids of prolactin hormones in various organisms, thicker lines showing higher incidences of amino acids

nitrogen count, weight of proteins and the hydrogen count were high, the support still was 89%. When the values for the carbon count, half-life of yeast and half-life of *E. coli* were high, the supports were 88%. Full details of all rules have been presented in Table 1.

Tree induction

When decision tree and ID3 models with gain ratio, information gain, gini index or accuracy criteria were run, a simple decision tree with a depth of 1 and total accuracy of 18.03% ± 5.77%, 14.47% ± 6.10%, 17.20% ± 6.46% or 15.38% ± 8.24%, respectively were generated. The most important feature used to build the tree in all models was the count of Cys-Ile.

When decision tree was run on numerical variables with gain ration, information gain, gini index and accuracy criteria, trees with depths of 6, 9, 11 and 10, total accuracy of 35.56% ± 11.37%, 25.83% ± 8.83%, 12.17% ± 12.17% and 30.69% ± 9.97% and precision 60.00% ± 10.81%, 75% ± 11.86, 75% ± 10.53% and 100% were created, respectively. The most important features used to build the trees were the frequency of His-Ile, weight, and the frequency of nitrogen and the Asp-Lys count. A full detail of this tree has been presented in Fig. 2.

ID3 (on numerical variables) generated a decision tree with a depth of 13 and accuracy of 29.68% ± 18.03% and class precision of 75%, when criterion were set to gain ratio. The most important feature used to build the tree was the frequency of Leu-Arg. When it was run on information gain criterion, a tree with a depth of 9, accuracy of 27.22% ± 16.73% and class precision of 71.42% was generated. Protein weight was the most important attribute to build the tree. Gini index criterion created a complex tree with a depth of 12 and accuracy of 25.83% ± 11.29% and class precision of 100%. The frequency of nitrogen was the main attribute. A complex tree with a depth of 45 and accuracy of 21.11% ± 14.35% and class precision of 100% was generated when ID3 criterion was set to accuracy and N-terminal amino acid was chosen as the most important protein feature.

C5.0 model generated a decision tree with a depth of 13 and cross-validation of 33.2 ± 3.7. The most important feature used to build the tree was the frequency of Arg-Arg.

When C&RT node was run on numerical data, the tree with a depth of 5 based on the frequency of Gln was generated. A tree with the same depth created with the Quest model and the frequency of Gly-Met was the most important feature.

When the CHAID model was applied to the data with and without feature selection, a tree with a depth of 5 was generated on the frequency of Ala-Arg attribute. The Asp count, the frequency of Phe-His and the Trp count were the most important protein features used by Simple Vote (Random Forest), decision Stump and random tree

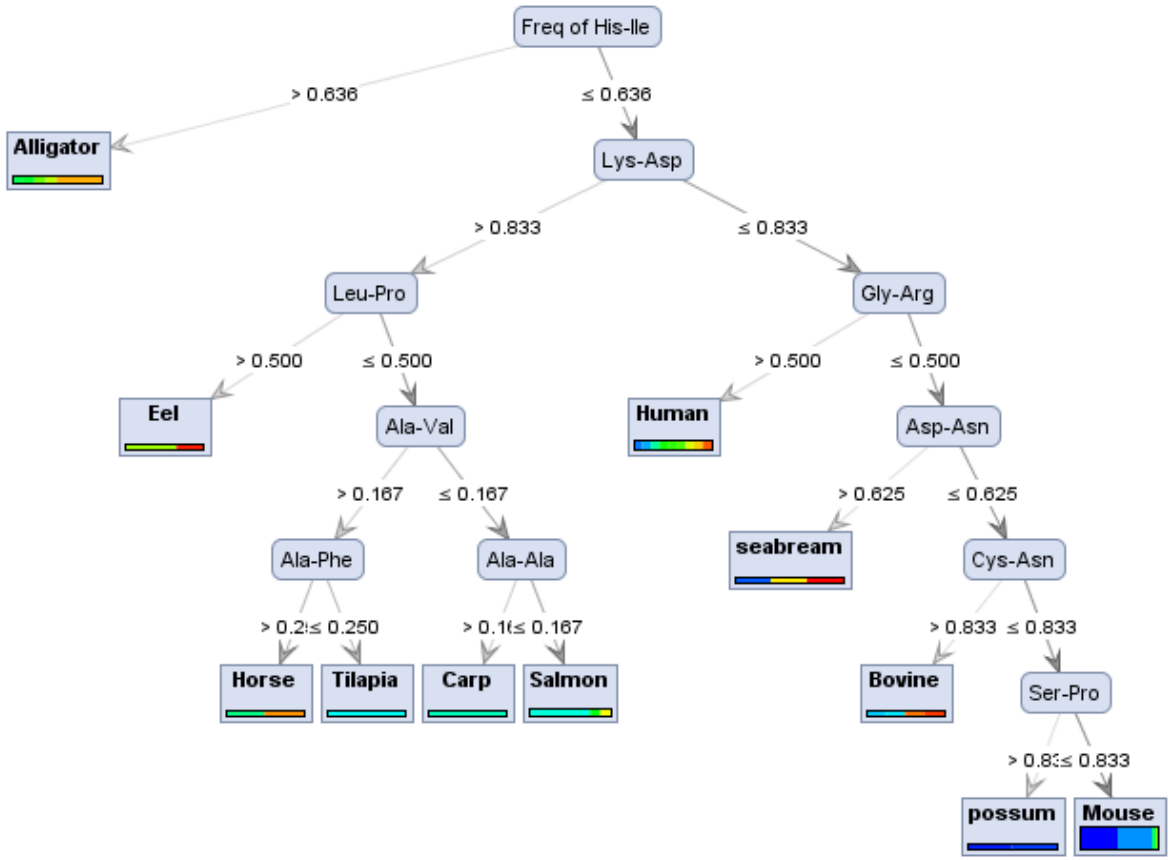

Fig. 2: Decision tree induced by decision tree algorithm (gain ratio criterion) run on numerical variables

algorithms to build the trees.

Rule induction

The frequency of nitrogen and the frequency of Ile-Met were selected as the most important features when rule induction (with information gain and accuracy criteria) were run on discretized dataset. But when the same models were run on numerical data, the frequency of nitrogen and the Gln count were the most important feature to build the rules (Table 1).

GRI node analysis created 100 rules with 85 valid transactions with minimum and maximum support of 12.94 and 17.65%, respectively. Maximum confidence reached 100% and minimum confidence decreased to 93.33%. When feature selection was used, minimum support, maximum support, maximum confidence, and minimum confidence changed to 11.67, 22.35, 100, and 84.21%, respectively. The highest confidence in without feature selection filtering occurred when the Gly count was less than 0.032, the frequency of Asp-Gln was less than 0.002 and the frequency of Lys-Ala was less than 0.004; while the same confidence with feature selection modeling occurred when the frequency of Ser-Arg, the frequency of Val-Ala and the frequency of Glu-Ala were less than 0.002, 0.004 and 0.008, respectively.

Clustering models

In K-Means model, 19 records were put into the first cluster and 22, 31, 10, and 3 records were put into the second, third, fourth, and fifth clusters, respectively, with a starting iteration of 6.23. When the K-Means model was applied on the dataset with feature selection filtering, again five clusters (with a starting iteration of 3.19) were generated, with 35, 19, 20, 5, and 6 recorded in each cluster, respectively.

Two Step cluster model clustered records into two groups with 35 and 50 records in each cluster, respectively. Only two clusters (with 19 and 66 records in each cluster) were created for the dataset filtered using feature selection criteria.

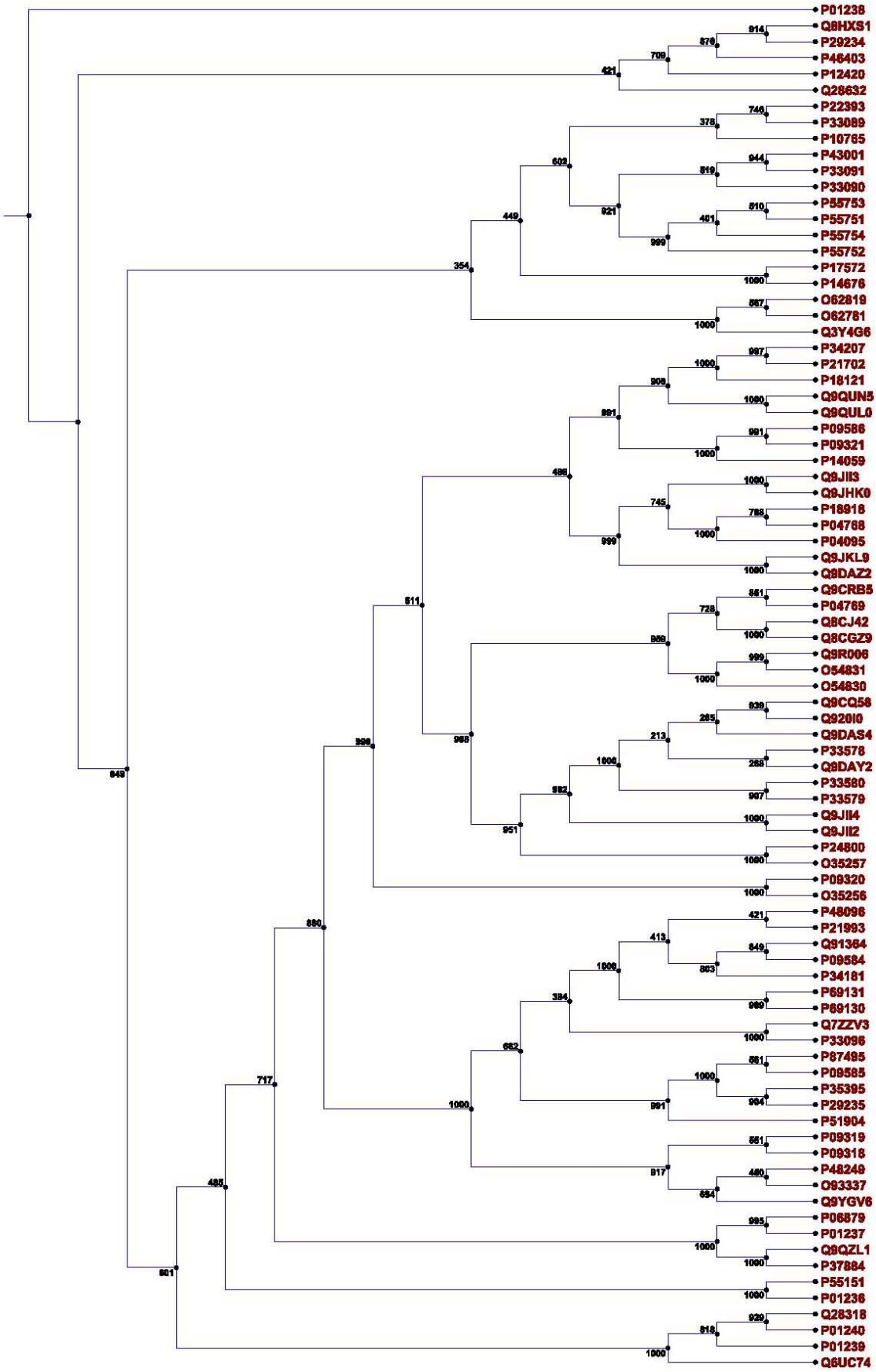

Fig. 3: Phylogenic tree of various prolactin hormones studied in the paper. The numbers represent prolactin protein accession number

A phylogenic tree for all prolactin hormone sequences generated by MEGA 4 Software, shown in Fig. 3.

Discussion

Here we extracted FASTA sequences of all reviewed prolactin hormones from Swiss-Prot protein databank. The reviewed proteins are those sequences which have been reviewed and verified by experts so they are original and not duplicated. Although the length and the weight of prolactin hormones were similar, the modeling applied here showed their protein attributes are not similar and modeling techniques can be used to categorize them. This is an important point which underlines previous assumptions that prolactin, growth hormone and somatolactin, together with the mammalian placental lactogen constitute a gene family of hormones with similar gene structure. These hormones are believed to have evolved from a common ancestral gene through several rounds of gene duplication and subsequent divergence (Huang *et al*., 2009). The results of this study showed even prolactin hormones may not share exactly all protein features, and this point can be used as a pivotal point to concentrate on for more investigation in future.

Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific research (Pham, 2008). The widening gap between data and information calls for a systemic development of data mining tools that will turn data tombs into golden nuggets of knowledge. To date, various methods have been employed to study the protein features in various animals (see introduction); here we applied different modeling techniques (screening, clustering, and item set mining and decision tree) to study more than 899 features of prolactin hormones in an attempt to determine species-related protein attributes.

The numbers of important protein attributes determined by attribute weighting models (weight higher than 0.07) were different as each model uses special statistical procedures to determine the most important features. Comparing the models' accuracy and performance, the best figure is obtained by gain ratio model. As shown in the results section, the frequency and the count of dipeptides play major roles in attribute weighting models; confirming the importance of dipeptides bonds in forming the proteins (Abdelmagid and Too, 2008). The frequency of Lys-Asp and the frequency of Phe-His were chosen by 60% of attribute weighting algorithms and the frequency of Ser-Asp, the count of Lys-Asp, the frequency of Leu-Arg, the frequency of Pro-Met, the count of Glu and the count of Phe-His were selected by 50% of attribute weighting models as important features.

Item set mining is a data analysis method that was originally developed for market basket analysis (Rotter *et al*., 2010). It aims at finding regularities in the shopping behavior of the customers of supermarkets, mail-order companies and online shops. In particular, it tries to identify sets of products that are frequently bought together. Once identified, such sets of associated products may be exploited to optimize the organization of the products on the shelves of a supermarket or on the pages of a mail-order catalog or web shop, may be used to suggest other products a customer could be interested in, or may give hints which products may be conveniently bundled. As a powerful method, it was quickly employed as a suitable tool in other data mining applications as well as bioinformatics modeling. It is acceptable in item set mining algorithms to have a large number of rules which should be refined and trimmed to just a few important rules as done here. The type of data was changed to binominal as requested by the model, setting the criteria to low, mid and high (>0.3, between 0.3 and 0.5 and higher than 0.5, respectively). Therefore, for each attribute three possible conditions (low, mid or high) were available and, at any time, just one condition will be true and two will be false. The numbers of low attributes were dominant in the dataset as the frequencies of attributes were generally less than 0.3; so to remove this dominant effect, we discard the low value in another attempt to see the real effects of mid and high values. Although the numbers of rules decreased to just 123 rules, the most important features forming the rules were

similar, showing and confirming the importance of data cleaning to reduce the burden on processing facilities.

Various decision tree models were performed on datasets with and without feature selection criteria. In some of them (such as ID3) a very simple tree was created with just one attribute and the count of Cys-Ile was used to build the tree, while in other models trees with more branches and depths were created. It would be reasonable to see some simple decision trees as prolactin hormones come from an ancestral gene, and so are in a similar rank. But in some decision tree models such as decision tree algorithm, the computation criteria were different and more complicated trees were generated. In all tree induction models the frequencies or the counts of dipeptides were the main protein attributes to build the tree, although the numbers and the types were different. The nitrogen count and weight of prolactin hormones in a few models were selected as the most important protein features. To our knowledge, this is the first study in this field to employ tree induction models to determine the specific structural properties of hormones and no such data was found on prolactin hormones. The findings confirmed that structural features of proteins, and especially hormones, may be used as important features in classifying hormones and opens up a new vista in this filed.

The results showed that various bioinformatics tools and modeling facilities can be used to identify the species specificity of prolactin hormones in animals with an acceptable precision rate. To our knowledge, for the first time we have shown that structural features such as primary or secondary protein attributes play an important role in prolactin hormones classification. This is also the first report on the importance of dipeptides' counts and frequencies in prolactin classification, but recently a few reports have emerged showing the importance of these features in other proteins' clustering (Ebrahimi *et al*., 2009; Bijanzadeh *et al*., 2010; Ebrahimi and Ebrahimie, 2010). It has also been reported here that feature selection or attribute weighting can be used to select the most important protein attributes and to reduce the burden of processing. The new findings open up new windows in understanding the characters' of prolactin hormones and also paves the way to engineer more efficient hormones in the lab by using various mutagenesis tools such as site directed mutagenesis.

Acknowledgements

The author greatly appreciates and acknowledges the support of Bioinformatics Research Groups, Green Research Center, and the School of Basic Sciences, Qom University for supporting the project.

References

Abdelmagid, SA and Too, CK (2008). Prolactin and estrogen up-regulate carboxypeptidase-d to promote nitric oxide production and survival of mcf-7 breast cancer cells. Endocrinology. 149: 4821-4828.

Ben-Jonathan, N; LaPensee, CR and LaPensee, EW (2008). What can we learn from rodents about prolactin in humans? Endocr. Rev., 29: 1-41.

Bijanzadeh, E; Emam, Y and Ebrahimie, E (2010). Determining the most important features contributing to wheat grain yield using supervised feature selection model. Aust. J. Crop. Sci., 4: 402-407.

Dancey, D; Bandar, ZA and McLean, D (2007). Logistic model tree extraction from artificial neural networks. IEEE Trans. Syst. Man. Cybern. B. Cybern., 37: 794-802.

Ebrahimi, M and Ebrahimie, E (2010). Sequence-based prediction of enzyme thermostability through bioinformatics algorithms. Curr. Bioinf., 5: 195-203.

Ebrahimi, M; Ebrahimi, E and Ebrahimi, M (2009). Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree algorithms. EXCLI J., 8: 218-233.

Gromiha, MM (2007). Prediction of protein stability upon point mutations. Biochem. Soc. Trans., 35: 1569-1573.

Gromiha, MM and Yabuki, Y (2008). Functional discrimination of membrane proteins using machine learning techniques. BMC Bioinf., 9: 135.

Hiyama, G; Sato, T; Zadworny, D and Kansaku, N (2009). Cloning of PRL and VIP cDNAs of the Java sparrow (*Padda oryzivora*).

Anim. Sci. J., 80: 176-186.

Huang, X; Hui, MN; Liu, Y; Yuen, DS; Zhang, Y; Chan, WY; Lin, HR; Cheng, SH and Cheng, CH (2009). Discovery of a novel prolactin in non-mammalian vertebrates: evolutionary perspectives and its involvement in teleost retina development. PLoS One. 4: 6163.

LaPensee, EW; Schwemberger, SJ; LaPensee, CR; Bahassi el, M; Afton, SE and Ben-Jonathan, N (2009). Prolactin confers resistance against cisplatin in breast cancer cells by activating glutathione-S-transferase. Carcinogenesis. 30: 1298-1304.

Liu, J; Xu, W; Sun, T; Wang, F; Puscheck, E; Brigstock, D; Wang, QT; Davis, R and Rappolee, DA (2009). Hyperosmolar stress induces global mRNA responses in placental trophoblast stem cells that emulate early post-implantation differentiation. Placenta. 30: 66-73.

Nichols, JJ and Green-Church, KB (2009). Mass spectrometry-based proteomic analyses in contact lens-related dry eye. Cornea. 28: 1109-1117.

Pariante, CM (2008). Pituitary volume in psychosis: the first review of the evidence. J. Psychopharmacol., 22: 76-81.

Pham, TD (2008). Computational prediction models for cancer classification using mass spectrometry data. Int. J. Data Min. Bioinform., 2: 405-422.

Rotter, A; Novak, PK; Baebler, S; Toplak, N; Blejec, A; Lavrac, N and Gruden, K (2010). Gene expression data analysis using closed item set mining for labeled data. OMICS. 14: 177-186.

Swaminathan, G; Varghese, B and Fuchs, SY (2008). Regulation of prolactin receptor levels and activity in breast cancer. J. Mammary Gland Biol. Neoplasia. 13: 81-91.

Trott, JF; Vonderhaar, BK and Hovey, RC (2008). Historical perspectives of prolactin and growth hormone as mammogens, lactogens and galactagogues-agog for the future. Neoplasia. 13: 3-11.

Ye, X; Fu, Z; Wang, H; Du, W; Wang, R; Sun, Y; Gao, Q and He, J (2009). A computerized system for signal detection in spontaneous reporting system of Shanghai China. Pharmacoepidemiol. Drug Saf., 18: 154-158.

Zhu, L; Yang, J; Song, JN; Chou, KC and Shen, HB (2010). Improving the accuracy of predicting disulfide connectivity by feature selection. J. Comput. Chem., 31: 1478-1485.