Title
Bioinformatic Analyses of the KX Blood-Group Antigen (KXA) Family and its inclusion in the Transporter-Opsin-G Protein-coupled Receptor (TOG) Superfamily

Permalink
https://escholarship.org/uc/item/1h98k9k7

Author
Contreras, Victor Manuel

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Bioinformatic Analyses of the KX Blood-Group Antigen (KXA) Family and its inclusion in the Transporter-Opsin-G Protein-coupled Receptor (TOG) Superfamily

A Thesis submitted in partial satisfaction of the requirements for the Degree Master of Science

in

Biology

by

Victor Manuel Contreras

Committee in charge:

Milton Saier, Jr., Chair
Aaron Coleman
Li-fan Lu

2016

The Thesis of Victor Manuel Contreras is approved and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2016

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Saier, for his guidance and support while working to complete this project. I would also like to thank all the fellow researchers that I have worked with in the Saier laboratory.

ABSTRACT OF THE THESIS

Bioinformatic Analyses of the KX Blood-Group Antigen (KXA) Family and its inclusion into the Transporter-Opsin-G Protein-coupled Receptor (TOG) Superfamily

by

Victor Manuel Contreras

Master of Science in Biology

University of California, San Diego, 2016

Professor Milton Saier, Jr., Chair

A bioinformatic characterization of the KX Blood- Group Antigen (KXA) Family has been conducted. This family includes the XK-related protein 8 which has recently been shown to function as a phospholipid flippase (Suzuki et al, 2013). This however, may not be its primary function. Multiple alignments, phylogenetic trees, and sequence analyses allowed me to characterize the KXA Family, revealing 17 primary clusters. Also, using average hydropathy, amphipathicity, and similarity plots, I was able to show that the members of this family possess on average 9 transmembrane helices (TMSs). Although, previous work has mostly been done with animal proteins, I demonstrate that homologs can be found in plants and ciliates. Furthermore, I show that the KXA family is a member of an established superfamily (TOG) of proteins having transport, light reception, electron flow, and G-protein receptor functions.

INTRODUCTION

The Kell blood group system found in erythrocytes is composed of two distinct membrane bound proteins XK and Kell and is the third most important blood group after the ABO and Rhesus systems to be considered before blood transfusions [9]. Kell is a 732 amino acid (aa) type II glycoprotein with a single transmembrane domain [12] and has been shown to contain all but one of the antigens that are attributed to this blood group system. Previous studies have shown that Kell belongs to the M13 family of endopeptidases and preferentially cleaves big ET-3 into ET-3 [13]. Although endothelins are known to have vasoconstrictive effects they are also involved in the endothelin-mediated release of nitric oxide, a potent vasodilator.

Cysteine residues in positions 72 and 347 in Kell and XK, respectively, connect the two proteins together via a disulfide bridge [23]. While the XK and Kell proteins form a heterodimer on the surface of red blood cells, in non-hematopoietic cells this heterodimer does not exist, and often the two proteins are not found within the same tissue [6, 14]. XK can be found ubiquitously throughout the body [14] where it may play different roles. The Kx antigen is found within the XK protein, and its absence along with a weak expression of the Kell antigens is characteristic of the McLeod syndrome [1]. The McLeod syndrome (MLS) is an X-linked disorder that is caused by a mutation in the XK gene which encodes the XK protein. Some of the common features of MLS include acanthocystis, haemolysis, cardiomyopathy, choreatic movement disorder, and seizures [1, 4, 33]. Recent studies have demonstrated that Xkr8, an animal member of the XK family functions as a phosphatidylserine (PS) exporter after being cleaved by caspase-3 during apoptosis. The accumulation of PS on the surface of the dying cell then serves as an “eat-me” signal to circulating macrophages [30].

In the present study we provide evidence for a relationship between members of the SWEET family of sugar transporters and XK-related proteins. SWEET and semi-SWEET proteins are members of the large and diverse Transporter-Opsin-G Protein-coupled Receptor (TOG) Superfamily [37] and are found ubiquitously and function as sugar exporters [3, 36, 27]. In plants, SWEETs function to export mono-saccharides and di-saccharides, playing a critical role in pollen secretion [7, 29] and nectar secretion [14]. In mammals, they may play a key role in allowing the liver to maintain appropriate glucose levels [32].

Bioinformatics analyses have shown that all members of the TOG superfamily arose by a duplication of a 4 TMS encoding genetic element to give 8 TMS proteins, although most members of the superfamily have lost a TMS, either the N-terminal TMS (most frequent) or the C-terminal TMS (relatively rare) [26,37].

COMPUTATIONAL METHODS

The HHpred server [28] was used to screen the protein Databank (PDB) using *Mus musculus* XK-related protein 8 (GI# 41235770) and other members of the XK family. Protein sequence alignments were created using ClustalX version 2.0 [11]. The FigTree program version 1.4.2 (created by the Rambaut laboratory) was used to generate phylogenetic trees. Protein average hydropathy, amphipathicity and similarity analysis was performed by usage of the AveHas program [38]. To determine distant homology between two distinct proteins, laboratory protocols were performed using the BioV suite [21]. Protocol 1 was used to perform PSI-BLAST searches on the National Center for Biotechnology (NCBI) non-redundant protein database for the two proteins in question. This search generates two lists of homologs for each original query protein presented as FASTA files. Alignments (using the Smith Waterman algorithm) were then performed between all FASTA files generated from the two different original query proteins using Protocol 2. The result of each alignment is then assigned a comparison score expressed in standard deviations (S.D.). Results from Protocol 2 were then re-aligned using the Global Sequence Alignment Tool (GSAT) which performs a Needleman Wunsch global alignment. Settings were established at 20,000 random shuffles; and more precise scores were once again determined. An alignment of 60 residues or more with a comparison score of 15 S.D. is deemed sufficient to establish homology between the original query proteins and the protein families from which they originated.

Chapter 1: Bioinformatic analyses of the KX Blood-group Antigen (KXA) Family

2.A.112: The KX Blood-group Antigen (KXA) Family

The KX Blood-group Antigen (KXA) Family, which is classified as the 2.A.112 Family in the Transporter Classification Database (TCDB) is a collection of proteins related to the membrane transport protein XK. Previous work [30] demonstrated that a member of this family, XK-related protein 8, can function as a phospholipid flippase to relocate phosphatidylserine from the inner leaflet of the cell membrane to the outer leaflet of the same membrane and serve as a signal for apoptosis. XK however, is ubiquitously present throughout the body [14], and its function has been theorized to be tissue-dependent. While homologs to XK-related proteins are currently thought to be found only in animals, this study demonstrates that homologues can be found in plants and ciliates as well.

1.1: Establishing subfamilies within the KX Blood-group Antigen (KXA) Family

Characterization of the KXA family began with closer inspection of one of the best understood members, the *Mus musculus* derived XK-related protein 8 (XKr8) (Uniprot ID: Q8C0T0), which has been shown to function as a phospholipid flippase [30]. This protein was added to the transporter classification database (TCDB) and given TC# of 2.A.112.1.3. By using the PSI-BLAST search tool within TCDB (TC BLAST), proteins closely related to XKr8 were identified and their TC numbers were reassigned to reflect their relationship with this protein. Further TC BLAST analysis of 2.A.112.1.3 and its related proteins found a series of potential distantly related homologs. Four main groups were identified into which all of these potentially homologous proteins fell. An analysis of members of each of these groups using the CDD program revealed that each possessed a protein segment found to be homologous to the KXA family. This finding led to the theory that these four groups represent four potential subfamilies within the larger KXA family. They were thus categorized as subfamilies one, two, three and four. Representative members of each subfamily also received a new TC number that reflected the subfamily to which it belonged, illustrating its overall relationship to the other members of KXA family.

1.2: Proving homology between the KX Blood-group Antigen (KXA) subfamilies

PSI-BLAST and CDD analyses suggested homology between the four different subfamilies within the KXA family. This suggestion was supported by additional statistical analyses (Table 1)

Subfamily 1 (TC# 2.A.112.1) and subfamily 2 (TC# 2.A.112.2)

The top comparison analysis between subfamily 1 and subfamily 2 homologues, Bfl4 (C3ZRW5) v Csu1 (I0Z719), gave a comparison score of 17 S.D. TMSs 2-6 of each of these proteins aligned with each other (Figure 1). A score of 46 S.D was obtained when comparing the complete sequences of Bfl4 and an established subfamily 1 member, TC# 2.A.112.1.3. In addition, a score of 94 S.D. resulted from the comparison of the full sequences of Csu1with a subfamily 2 member, TC# 2.A.112.2.1. These characteristics meet our statistical criteria for homology, and thus using the superfamily principle, it could be confirmed that these two subfamilies are homologous.

Subfamily 1 (TC# 2.A.112.1) and subfamily 3 (TC# 2.A.112.3)

The last 8 TMSs of subfamily 1 homologue, Lan1 (XP_013411986.1) and subfamily 3 homologue, Imu1 (G0QX98) aligned with a score of 13 S.D. (Figure 2). Despite the high number of TMSs aligning between homologues of subfamily 1 and subfamily 3 an alignment score of less than 15 S.D. does not allow us to make confident statistical predictions on the relationship of these two subfamilies. Additional evidence will be presented later in this study to establish homology between subfamily 1 and subfamily 3.

Subfamily 1 (TC# 2.A.112.1) and subfamily 4 (TC# 2.A.112.4)

A comparison between the subfamily1 homologue Zne1 (A0A067R5K1) and subfamily 4 homologue Bmo1 (XP_012548325.1) resulted in an alignment of the first 5 TMSs of the two proteins with a score of 21 S.D. (Figure 3). A score of 27 S.D. was obtained when comparing the full sequence of Zne1with the full sequence of a member of of subfamily 1, TC# 2.A.112.1.3. Similarly, a score of 63 S.D. resulted from the full length comparison between Bmo1 and a member of subfamily 4, TC# 2.A.112.4.1. These high scores fulfill our established requirements for proving homology between protein sequences. Thus, it was confirmed that subfamilies 1 and 4 of the KXA family are homologous to each other.

Subfamily 2 (TC# 2.A.112.2) and subfamily 3 (TC# 2.A.112.3)

The first 8 TMSs of subfamily 2 homologue Spu2 (W4Y075) and subfamily 3 homologue Tth16 (I7MG83) were compared, giving a score of 16 S.D. (Figure 4). The comparison of the full sequence of Spu2 with the full sequence of subfamily 2 member TC# 2.A.112.2.2 (K8EHG2) resulted in a score of 9 S.D. In addition, the full sequence comparison of Tth16 with the full sequence of subfamily 3 member TC# 2.A.112.3.2 yielded a score of 46 S.D. Since the comparison between Spu2 (W4Y075) and 2.A.112.2.2 gave a score lower than 15, further analysis was necessary to determine their relationship (Table 2). An alignment between the 2.A.112.2.2 homologue, Csu1 (I0Z719) and the W4Y075 homologue, Pre3 (XP_008426645.1) resulted in a score of 16 S.D. with 3 TMSs aligning. The full sequence comparison between Csu1(I0Z719) and 2.A.112.2.2 gave a score of 148 S.D. While the full sequence comparison of Pre3 (XP_008426645.1) and W4Y075 yielded in a score of 126 S.D.

The comparison scores between true homologues of Spu2 (W4Y075) and TC# 2.A.112.2.2 gave scores higher than 15 S.D. demonstrating that these two proteins are homologous, and this in addition to our previous results, allowed us to make the conclusion that subfamilies 2 and 3 within the KXA family are homologous. Lastly, although the direct comparison of subfamilies 1 and 3 did not yield significant scores, homology between subfamilies 1 and 2, and homology between subfamilies 2 and 3 has now been established. Therefore, the superfamily principle can once again be applied to confirm that subfamilies 1 and 3 are homologous.

Subfamily 2 (TC# 2.A.112.2) and subfamily 4 (TC# 2.A.112.4)

The comparison between the subfamily 2 homologue Smi2 (A0A087UTH8) and subfamily 4 homologue Mde1 (XP_008544814.1) aligned the first 5 TMSs of these two proteins with a score of 21 S.D. (Figure 5). The full sequence comparison of Smi2 and subfamily 2 member TC# 2.A.112.2.1 (A4RVV9) gave a score of 6 S.D. While the full length comparison between Mde1 and subfamily 4 member TC# 2.A.112.4.1 (B4J9J5) gave a score of 61. Since the comparison between Smi2 (A0A087UTH8) and TC# 2.A.112.2.1 (A4RVV9) did not yield significant results, further analysis was necessary to establish their relationship (Table 2). An alignment between the A4RVV9 homologue, Csu1 (I0Z719) and the A0A087UTH8 homologue, Bfl4 (C3ZRW5) yielded 4 TMSs aligning with a score of 17 S.D. A comparison of the full sequences of A4RVV9 and I0Z719 gave a score of 115 S.D., while a comparison of A0A087UTH8 and C3ZRW5 yielded a score of 21 S.D. These results meet our criteria for establishing homology not only between Smi2 (A0A087UTH8) and TC# 2.A.112.2.1 (A4RVV9), but also for subfamily 2 and subfamily 4 of the KXA family.

Subfamily 3 (TC# 2.A.112.3) and subfamily 4 (TC# 2.A.112.4)

TMSs 4-5 of the subfamily 3 homologue Tth1 (I7MG83) and TMSs 4-5 of the subfamily 4 homologue Ppo1 (XP_013133483.1) aligned with a score of 15 S.D. (Figure 6). Comparisons of the full length sequences of Tth1 and the subfamily 3 member TC# 2.A.112.3.1 (I7MG83) gave a score of 266 S.D. In addition, the full sequence comparison between Ppo1 and subfamily 4 member TC# 2.A.112.4.1 (B4J9J5) yielded a score of 70 S.D. All these results confirm that subfamily 3 and subfamily 4 are homologous.

1.3: Phylogenetic Analysis

After homology was established between all subfamilies of the KX Blood-group Antigen (KXA) family, representatives from each subfamily were taken and used as query proteins for BLAST searches that would identify and generate a list of members of the family. An 80% cutoff was used to eliminate redundant proteins. A Clustal X alignment was performed with the resulting proteins and was then used to create a phylogenetic tree (Figure 7). The following is an analysis of the results gathered for the different protein clusters in the phylogenetic tree.

A total of 248 KXA Family members were analyzed in this study. These were divided into 17 independent clusters based on their position in the phylogenetic tree (Figure 7). When visible sub-clusters were present within any given cluster, these sub-clusters were labeled with an appropriate letter. The proteins were listed in Table 3 as they appeared in clockwise fashion around the phylogenetic tree, starting with Cluster 1A.

Cluster 1A (16 proteins) is derived from different species of fish. Considering the tightness within this particular sub-cluster, it is no surprise that the proteins found within it are so closely related. Further analysis via BLAST indicted that all the proteins in this sub-cluster are similar to XK-related protein 9. The average size of these proteins is 572 +/- 20 amino acids (aas). The number of predicted transmembrane segments (TMSs) varied across the sub-cluster (the highest number of predicted TMSs was 9 and the lowest number was 6) with an average of 8 TMSs overall. It is interesting to note that the first 6 proteins in this sub-clusters (Ssa3, Elu4, Cha11, Sfo2, Dre3, and Ame6) cluster more closely together and are on average smaller than proteins 7-16.

Cluster 1B (28 proteins) is similar in size to cluster 1A. The average size of these proteins is 377 +/- 7 aas. Little variation in the number of predicted TMSs across each protein was found with most members having 9 TMSs with a range from 7 to 10. Cluster 1B proteins derive from many different species of mammals, reptiles, birds, and fish. Despite this organismal diversity, these proteins are also homologous to XK-related protein 9.

Cluster 2A (12 proteins) derives from various species of fish. The average size of these proteins is 414 +/- 10 aas, significantly larger than the two clusters previously described. The average number of TMSs found is 9, with a range from 8 (Ame1) to 10 (Cse3 and Cmi1). BLAST analysis showed that these proteins are most closely related to XK-related protein 8.

Cluster 2B (16 proteins) exhibits proteins with an average size of 406 +/- 17 aas They derive from species of mammals, birds, and reptiles. Proteins Ete3, Hsa3, Mdo5, Gga2, Zal1, and Ami3 are more distantly related than the other proteins found in this sub-cluster. Ete3 is the largest protein found in cluster 2B with 451 aas, most of it consisting of an XK-related domain. Although the average number of TMSs in this cluster is 8, there was greater variability than in the previous clusters. The range was from 5 to 10 TMSs. CDD analysis for Hsa3 revealed one XK-related domain from residues 190 to 350. All proteins within cluster 2B are closely related to XK-related protein 8.

Cluster 2C (18 proteins) is similar to protein clusters 2A and 2B. The average size of the proteins in cluster 2C is 416 +/- 32. aas. The average number of TMSs is 9, but there is great variety for the individual proteins, with the highest number being 11 and the lowest being 7. Protein Pfo2 is most distantly related to the other proteins in this cluster and is the largest protein with 535 aas and 11 TMSs. A search for functional domains with CDD revealed an XK-related segment from residues 110 to 450. According to HMMTOP, the last predicted TMS within Pfo2 is not part of the detected functional domain. If we exclude Pfo2, the average protein size is 409 aas +/- 12 aas. All proteins within this clusters were found to be closely related to XK-related protein 8.

Cluster 3A (6 proteins) proteins are on average larger than the proteins found within the previous clusters. The average protein size is 690 +/- 28 aas. The average number of predicted TMSs is 6, although the variability is considerable. Sha4 and Mdo4 are separated from the other members both phylogenetically and in size, being larger and with more TMSs than the other members of cluster 3A. CDD analysis showed that the XK-related functional domain within Sha4 is from residue 20 to residue 320, while the same functional domain was found in Mdo4 from residues 5 to 290. The last TMS of Mdo4 is outside of the the functional domain.

Cluster 3B (3 proteins) is derived from species of reptiles and birds. The average size of these proteins is 668 +/- 62 aas. The high standard deviation value is due to the large size of Cpi2, which has a length of 737 aas. The XK functional domain is between residues 20 and 320 in Cpi2, and this domain includes all TMSs. The average TMS number in the cluster is 8. All the proteins in this cluster showed greatest similarity to XK-related protein 5 in TCDB.

Cluster 3C (6 proteins) derives from fish species. These proteins have an average size of 649 +/- 29 aas with considerable variability. Likewise, although the TMS average is 9, the range is 7 to 11. Surprisingly, the smallest protein, Loc5, is the one with the largest number of predicted TMSs. Loc5 has the XK related domain from residues 20 to 325 with the TMSs within this functional domain. All proteins within cluster 3C showed strongest similarity to XK-related protein 5 in TCDB.

Cluster 4A (5 proteins) is derived from different species of fruit flies. These proteins have an average size 352 +/- 5 aas, with an average of 9 TMSs. All proteins in this cluster showed strong similarity to a protein derived from *Drosophila melanogaster,* TC# 2.A.112.1.5.

Cluster 4B (5 proteins) has an average protein size of 432 +/- 23 aas with 9 TMSs and derives exclusively from flies. These proteins exhibit greatest similarity to the same protein as for cluster 4A.

Cluster 5A (10 proteins) derives from different species of insects. These proteins have an average size of 406 +/- 21 aas with an average number of 9 predicted TMSs and exhibit the greatest similarity to a protein from *Daphnia pulex,* TC# 2.A.112.1.10. These proteins are closely related to XK-related protein 6.

Cluster 5B (15 proteins) derives from a wide variety of insects. There are only minimal differences in size across the entire cluster with the average length of 428 +/- 11 aas and an average of 9 predicted TMSs. All the proteins in this cluster demonstrated strong similarity to the same protein as for cluster 5A. These proteins were also closely related to XK-related protein 6.

Cluster 6 (4 proteins) is made up of proteins that exhibit an average size of 423 +/- 3 aas and possess 9 predicted TMSs. These proteins derive from different species of insects and exhibit XK-related functional domains where all predicted TMSs can be found. Like the proteins in clusters 5A and 5B, they resemble the protein derived from *Daphnia pulex,* TC# 2.A.112.1.10, as well as XK-related protein 6.

Cluster 7 (2 proteins) consists of proteins Smi1 and Moc2. Although both are from arachnids, their sizes and numbers of TMSs differ. Smi1 has a size of 455 aas with 9 predicted TMSs and a large XK-related functional domain (residues 88 to 440). This protein proved to have strong similarity to the protein derived from *Daphnia pulex,* TC# 2.A.112.1.10. Moc2, on the other hand, has a size of 563 aas, with 7 predicted TMSs and an XK-related functional domain from residues 170 to 545 with all detected TMSs within this particular domain. However, the hydropathy and amphipathicity plot generated by the WHAT program revealed two small hydropathic peaks not predicted to be TMSs. Unlike, Smi1, Moc2 proved to possess strongest similarity to a protein derived from *Branchiostoma floridae,* TC# 2.A.112.1.9. Both of these proteins however, resemble XK-related protein 6 in TCDB.

Cluster 8A (5 proteins) derives from fish. The proteins found in this cluster have an average size of 575 +/- 9 aas with 7-9 predicted TMSs. Cha3, the protein in this cluster with the most predicted TMSs has an XK-related functional domain from residues 80 to 435. All TMSs in Cha3 are located within this domain. All the proteins in this cluster showed greatest similarity to a human protein, TC# 2.A.112.1.11, found in TCDB. Despite the fact that this protein is classified as a XK-related 4 protein, NCBI BLAST analysis showed that all the proteins in cluster 8A are more closely related to XK-related protein 6.

Cluster 8B (7 proteins) consists of proteins from the animal kingdom, but they are highly diverse in size. These proteins have a average size of 636 +/- 129 aas, with the smallest (Pre2) having 400 aas and the largest (Fca1), 822 aas. When Fca1 is analyzed using CDD, an XK-related functional domain was detected from residues 280 to 680. When the last 200 residues were blasted against NCBI, many other XK-related proteins showed up, but when the same was done for the first 200 residues only Fca1shows high sequence similarity. Using the WHAT program, all 10 predicted TMSs of Fca1 were found within its XK-related functional domain. The majority of the proteins in this cluster also have 10 predicted TMSs. All proteins in this cluster demonstrated greatest similarity to the same human protein as the previous cluster, TC# 2.A.112.1.11. Most of the proteins in the cluster showed strongest similarity with XK-related protein 4, although Pre2 showed showed a stronger similarity with XK-related protein 7.

Cluster 8C (14 proteins) has an average size of 570 +/- 35 aas and is derived from different species of fish, reptiles and mammals. Although most proteins have 8 predicted TMSs, the range was 5-10. Xma1, the protein with the highest number of predicted TMSs, has an XK-related functional domain (residues 45 to 395) where all its TMSs are found. Proteins in this cluster demonstrated strong similarity to the same human protein as the 2 previous clusters, TC#

2.A.112.1.11 Despite, the fact that this protein is classified as a XK-related 4 protein, NCBI BLAST analysis suggested that cluster 8 proteins may be more similar to XK-related protein 7.

Cluster 9 (2 proteins) is composed of two sea urchin proteins (Spu2 and Spu3) with an average size of 551 +/- 1 aas. Both have 9 TMSs in an XK-related functional domain. All the proteins in this cluster demonstrated strong similarity to a protein derived from *Branchiostoma floridae* TC# 2.A.112.1.9. An NCBI BLAST search revealed that Spu2 was closely related to XK-related protein 6, while Spu3 was equally closely related to XK-related proteins 4 and 6.

Cluster 10 (9 proteins) derives from different species of worms, arachnids, and crabs. The proteins found in this cluster have an average size of 459 +/- 83 aas with 9 predicted TMSs on average. All predicted TMSs can be found within the XK-related domain found in these proteins. Differences in some proteins could be observed. Two proteins, Isc1 and Cte2, proved to be truncated. Another protein, Cte4 looked strikingly different from the rest of the proteins in this cluster, and had fewer hydrophobic peaks. Several proteins in this cluster are most similar to a protein derived from *Daphnia pulex,* TC# 2.A.112.1.10, but others are more closely related to a protein derived from *Branchiostoma floridae,* TC# 2.A.112.1.9. However all these proteins resemble XK-related proteins 4 and 6.

Cluster 11A (4 proteins) is derived from oysters and includes proteins that have an average size of 488 +/- 68 aas. Cgi1 was predicted to have 10 TMSs, Cgi5, 9 TMSs, and Cgi3 and Cgi8, 7 TMSs. Analysis of the hydropathy and amphipathicity plots revealed that Cgi1 and Cgi5 have similar hydropathic profiles; however Cgi5 has a medium sized hydrophobic peak that was not detected as a TMS by HMMTOP. The plots of Cgi3 and Cgi8 looked very different from each other and from those of the first two proteins as well. Cgi3 has 2 medium sized hydrophobic peaks not detected as TMSs by HMMTOP, while Cgi8 has only 1. A BLAST search against TCDB showed that Cgi1 and Cgi5 were most closely related to a protein in TCDB derived from *Branchiostoma floridae,* TC# 2.A.112.1.9. Cgi3 proved to be more closely related to a protein derived from *Daphnia pulex,* TC# 2.A.112.1.10. Lastly, Cgi8 was found to have greater similarity to to a human protein, TC# 2.A.112.1.11. All proteins in this cluster are related to XK-related proteins 4 and 6.

Cluster 11B (3 proteins) originates from 2 different species of molluscs. They have an average size of 548 +/- 55 aas and have 8-9 predicted TMSs. CDD analysis of the largest protein, Bgl3, found an XK- related functional domain from residues 200 to 545 with the predicted TMSs found within this domain. All proteins in this cluster were found to be most closely related (in TCDB) to a protein derived from *Branchiostoma floridae,* TC# 2.A.112.1.9, and similar to XK-related protein 6.

Cluster 12A (5 proteins) originates from different species of insects. The proteins found in this cluster have an average size of 586 +/- 8 aas with 9 TMSs each. Analysis of each protein within this cluster by CDD found 2 XK-related functional domains, one at each end of the protein. All the TMSs predicted in these proteins are found in either of these two functional domains. These proteins are closely related to a protein derived from *Aedes aegypti,* TC# 2.A.112.4.3.

Cluster 12B (6 proteins) is derived from species of flies. The proteins found in this cluster have an average size of 654 +/- 11 aas with 8-9 predicted TMSs. The proteins in this cluster are most closely related to a protein derived from *Drosophila grimshawi,* TC# 2.A.112.4.1.

Cluster 12C (5 proteins) is derived from different species of mosquito. It possesses proteins with an average size of 662 +/- 65 aas and 8 or 9 predicted TMSs. Analysis of each protein within this cluster by CDD found two XK-related functional domains, one at each end of the protein. Closer inspection of Ast1, the largest protein in this cluster, revealed the same two functional domains, with the only difference being that Ast1 had an extra segment of about 100 residues before the first domain. The proteins in these cluster are most closely related to a protein derived from *Aedes aegypti,* TC# 2.A.112.4.3.

Cluster 12D (14 proteins) originates from different species of insects and has proteins with an average size of 535 +/- 10 aas and 9 TMSs. CDD analysis detected different functional domains within these proteins. Eight of these proteins contain two XK-related functional domains, one close to the N termini and another closer to the C termini. However, proteins Cfl1 and Oab1 contain only one XK-related functional domain near the C terminus and as well as one Formin Homology Region 1 domain towards the middle of the protein. Protein Mro1 contains two identified XK-related functional domains, one on each terminus, the first containing TMSs 1-5 and the second containing TMSs 4-9, and one Formin Homology Region 1 in the hydrophilic middle section of the protein. Protein Tpr1 exhibited only one XK-related functional domain near the C terminus. Protein Mde1 was found to contain two XK-related functional domains, one close to its N terminus and the other close to its C terminus, as well as a central predicted coiled-coil domain (DUF2360). Protein Mph1 also contained two XK-related functional domains, at its two ends as well as a central Sine oculis-binding protein (SOBP) functional domain. These proteins are closely related to a protein derived from *Solenopsis invicta,* TC# 2.A.112.4.2.

Cluster 12E (4 proteins) derives from insects and has proteins with an average length of 494 aas +/- 18 aas and 9 predicted TMSs, except for Dpo2, which has 11 predicted TMSs. Analysis of the hydropathic and amphipathic plots of Dpo2 and the other proteins in this cluster revealed that Dpo2 has an extra hydropathic peak from residues 327-344 as well as another large peak that was predicted to include two TMSs instead of one. CDD analysis revealed two XK-related functional domains in Dpo2, one close to each terminus, and both predicted to contain all of its potential TMSs. These proteins are closely related to a protein derived from *Solenopsis invicta,* TC# 2.A.112.4.2.

Cluster 13 (3 proteins) derives from 2 species of algae and 1 species of plankton and contains proteins with an average size of 928 aas +/- 173 aas. In addition to the significant size differences, these proteins also vary in the number of predicted TMSs. Protein Olu1 was predicted to have 7 TMSs, Csu1 9 TMSs, and Bpr1 10 TMSs. Analysis of the hydropathy and amphipathicity plots of these 3 proteins revealed significant peaks in proteins Olu1 and Csu1 (1 and 2 peaks, respectively) that were not interpreted as TMSs by HMMTOP. Despite the great differences in size and numbers of predicted TMSs, all of these proteins posses an XK-related functional domain. These proteins have been entered into TCDB. Olu1 with TC# 2.A.112.2.1, Bpr1 with TC# 2.A.112.2.2, and Csu1 TC# 2.A.112.2.3.

Cluster 14 (2 proteins) is derived from two protozoans. Tth1 was predicted to have 7 TMSs, while Imu1 was predicted to have 9. These two proteins have different topologies. CDD analysis identified several functional domains in Tth1 which include an XK-related domain, an uncharacterized coiled-coil DUF342 domain, a chromosome segregation protein (SMC) domain, a Myosin tail 1 domain, and a chromosome segregation protein (PRK03918) domain. A similar CDD analysis of Imu1 revealed several functional domains including a chromosome segregation ATPase, two chromosome segregation protein domains SMC (GR02168) and (PRK01156) as well as a plant domain of unknown function (DUF869). Both of these proteins were put into TCDB. Protein Tth1 with TC# 2.A.112.3.1 and Imu1 with TC# 2.A.112.3.2.

Cluster 15 (6 proteins) derives from different species of insects. The proteins that form this cluster have an average size of 825 +/- 203 aas. The high variability in the sizes of these proteins correlates with their substantial divergence in the phylogenetic tree. Although the average number of predicted TMSs found in this cluster is 10, some proteins have as many as 11 TMSs and as few as 8 TMSs. Interestingly, the largest protein Ssc1 is also the one with fewest predicted TMSs. A CDD analysis of this protein revealed a functional domain related to XK from residues 185-564 and overlapping with a Major Facilitator Superfamily (MFS) family permease domain from residues 399-568. A comparison of the hydropathy and amphipathicity plots of this protein and Ppo2, another protein in this cluster (11 TMSs), revealed that the two proteins have very different hydropathy profiles. These proteins are closely related to a protein derived from *Culex quinquefasciatus*, TC# 2.A.112.1.4.

Cluster 16 (6 proteins) is derived from different species of round worms. These proteins exhibit an average protein size of 457 aas +/- 15 aas with most proteins having 8 predicted TMSs. These proteins are closely related to a protein derived from *Caenorhabditis elegans,* TC# 2.A.112.1.2. Interestingly, these proteins are most closely related to Cell death abnormality protein-8, a well known XK-related homologue present in *C. elegans* [30].

Cluster 17 (7 proteins) derives from different species of insects. The proteins in this cluster possess an average size of 1130 aas +/- 93 aas and an average of 10 predicted TMSs. Proteins Pxu2 and Bmo4 were predicted to have both XK and MFS domains while Cfl2 has a putative RhaT domain within its XK-related functional domain. Proteins Pxy3, Aro1, Oab2, and Zne4 in this cluster resemble a protein derived from *Branchiostoma floridae,* TC# 2.A.112.1.9, while proteins Pxu2, Bmo4, and Cfl2 are more closely related to a human protein with a TC# 2.A.112.1.11.

1.4 Hydropathy analyses

A Clustal X alignment was generated using the FASTA sequences of all members of the KX Blood-group Antigen (KXA) Family included in TCDB. This alignment was then used to create average hydropathy, amphipathicity, and similarity plots (Figure 8) using the AveHas program. These plots revealed 9 well conserved hydrophobic peaks, suggesting 9 well conserved TMSs in all or most of the KXA subfamilies.

Chapter 2: Integration of the KX Blood-group Antigen (KXA) Family into the Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily

2.1: Integration of the KX Blood-group Antigen (KXA) family into the Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily

The studies reported in this section (results summarized in Table 4) demonstrate that the last 4 TMSs in the KXA family are homologous to the 4 TMS repeat units in members of the TOG superfamily.

The KXA family (TC# 2.A.112) and the Lysosomal Cystine Transporter (LCT) family (TC# 2.A.43)

Proteins that belong to the LCT family derive from plants, animals, and fungi and have 7 TMSs on average. A particular human homolog, Cystinosin, is encoded by the CTNS gene. Disruptive mutations in this gene cause the improper transport of cystine out of lysosomes, a condition called Cystinosis [31].

A comparison between the KXA homologue Ame5 (W5K2Y7) and the LCT homologue Jcu3 (XP_012079100.1) demonstrated an alignment of the last 3 TMSs in both proteins (TMSs 7-9 in Ame5 and TMSs 4-6 in Jcu3) with a score of 15 S.D. (Figure 9). A score of 31 S.D. was obtained when comparing the full length sequences of Ame5 with KXA protein, TC# 2.A.112.1.3, and a score of 34 S.D. was obtained when comparing the full sequences of Jcu3 with LCT protein, TC# 2.A.43.2.2. These results satisfy our statistical criteria for homology, when applying the superfamily principle.

The KXA family (TC# 2.A.112) and the Ni^{2+}-Co^{2+} Transporter (NiCoT) Family (TC# 2.A.52)

Members of the NiCoT family are responsible for catalyzing the transport of Nickel and Cobalt. These proteins possess on average 7-8 TMSs and are present in Bacteria, Archaea, and Eukaryotes.

TMSs 6-7 of KXA homologue Cha3 (XP_012677872.1) aligned with TMSs 4-5 of NiCoT homologue Sco3 (A0A0E9NFW4) with a score of 15 S.D. (Figure 10). A full sequence comparison of Cha3 with KXA protein, TC# 2.A.112.1.3 yielded a score of 45 S.D. In addition, a score of 21 S.D. resulted when comparing the full sequence of Sco3 with the full sequence of NiCoT protein, TC# 2.A.52.2.2. These results demonstrate homology between the KXA family and the NiCoT family.

The KXA family (TC# 2.A.112) and the Phosphate: Na^+ Symporter (PNaS) Family (TC# 2.A.58)

The PNaS family is composed of sodium-dependent inorganic phosphate transport proteins. These proteins are derived from mammals, fish, and *C. elegans* and have on average 8-12 TMSs. The last TMSs (TMSs 7-9) of Spa1 (XP_008273837.1), a KXA homologue, aligned with the last TMSs (TMSs 9-11) of Gja1 (XP_015272244.1), a PNaS homologue, with a comparison score of 14 S.D. (Figure 11). The full sequence alignment of Spa1 with the KXA protein, TC# 2.A.112.1.1 yielded a score of 157 S.D. while the comparison of the full length sequence of Gja1 and PNaS protein, TC# 2.A.58.1.5 resulted in a score of 239 S.D. Although a score of 14 S.D. is not sufficient to demonstrate homology, this new information, together with our other analyses showing KXA family members to be homologous to other members of the TOG superfamily serve as a strong argument that the KXA family is also homologous to the PNaS family.

The KXA family (TC# 2.A.112) and the Organic Solute Transporter (OST) Family (TC# 2.A.82)

The OST family is composed of proteins that come together to form heterodimers which transport a variety of organic compounds. Subunit α has 7 predicted TMSs while subunit β has only 1. These proteins can be found in humans, mice and *Raja erinacea* [25, 34].

A comparison between the KXA homologue, Cfl2 (E2A6Z0) and OST homologue, Mve5 (KFH63302.1) yielded a score of 17 S.D. with TMSs 4-9 (last 6 TMSs) in Cfl2 aligning with TMSs 2-6 of Mve5 (Figure 12). It is interesting to note that TMS 4 in Cfl2 aligned with sections of TMSs 6 and 7 in Mve5. A full length comparison of Cfl2 with the KXA protein, TC# 2.A.112.1.3 gave a score of 51 S.D., and score of 60 S.D. was achieved when comparing the full sequences of Mve5 and the established OST protein, TC# 2.A.82.1.3. The statistical significance of these scores demonstrate the homologous relationships between the KXA family and the OST family, again showing that the KXA family is a member of the TOG superfamily.

KXA family (TC# 2.A.112) and 4-Toluene Sulfonate Uptake Permease (TSUP) Family (TC# 2.A.102)

The TSUP family members can be found in both prokaryotes and eukaryotes [26]. Although the function of this family has not been fully characterized, there is evidence that members of the TSUP family serve as transporters for sulfur-containing organic compounds [17, 18, 35]. The 8 TMSs commonly found in this family have been theorized to have arisen from an ancestral 4 TMSs repeat unit (Shlykov et al, 2012).

Comparing TMSs 1-3 of the KXA homologue Tpr1 (XP_014235795.1) with TMSs 1-3 of the TSUP homologue Rba6 yielded a score of 15 S.D. (Figure 13). In addition, a score of 30 S.D. was calculated from the full sequence comparison of Tpr1 with KXA protein, TC# 2.A.112.1.12, while comparing the full sequences of Rba6 and TSUP protein, TC# 2.A.102.4.6

resulted in a score of 66 S.D. These results satisfy our criteria showing that the KXA family is homologous to the TSUP family, an established member of the TOG superfamily.

KXA family (TC# 2.A.112) and Sweet; PQ-loop; Saliva; MtN3 (Sweet) Family (TC# 2.A.123)

Sweet family members are found in all domains of life and function as sugar exporters [3, 27, 36]. The semi-Sweets, the smallest of the Sweet homologues, can be found in bacteria and consist of a triple-helix-bundle. Other SWEET homologues are predicted to have 7 transmembrane helices. It has been theorized that these 7 transmembrane segments could have arisen from an ancient duplication of the original 3 helix bundle with the addition of an extra central transmembrane helix (TMS) in a 3+1+3 configuration [3].

A comparison of Eeu2 (XP_004074365.1), a KXA homologue and Nsy2 (XP_009783143.1), a Sweet homologue, resulted in a top score of 20 S.D. between the two last TMSs of these two proteins (Figure 14a). Such an unusually high score suggests a very strong homologous relationship between KXA and Sweet. A comparison score of 24 S.D. resulted from the full sequence alignment of Eeu2 and the KXA protein, TC# 2.A.112.1.3, and the full length comparison of Nsy2 and the established Sweet protein, TC# 2.A.123.1.17 resulted in a score of 37 S.D. The results from this statistical analysis exceed our criteria for homology. An additional comparison between Cte1(R7UG75), a KXA homologue and Tur3 (M7ZA79), a Sweet homologue resulted in the alignment of the last 4 TMSs of these two proteins with a score of 15 S.D. (Figure 14b). A full sequence alignment between Tur3 and Sweet protein, TC #2.A.123.1.17 yielded a score of 42 S.D., while a full sequence alignment between Cte1 and the KXA protein, TC # 2.A.112.1.3, resulted in a score of 6 S.D. A full sequence alignment between Cte1 and Aca2 (XP_005104591.1) gave a score of 18 S.D. while a full sequence alignment between KXA protein, TC# 2.A.112.1.3 and Aca2 gave a score of 39 S.D. showing that Cte1 is truly homologous to members of the KXA family. These two new comparisons between the

KXA family and the Sweet family, not only show the homologous relationship between the two, but also shows the strong similarity between their last 4 TMSs. This result provides compelling evidence that the KXA family is part of the TOG superfamily.

The KXA family (TC# 2.A.112) and the Ion-translocating Microbial Rhodopsin (MR) Family (TC# 3.E.1)

The MR family is composed of membrane proteins that catalyze the light-powered translocation of ions across microbial membranes. Members of this family include bacteriorhodopsin [16], sensory rhodopsin II, [22], and halorhodopsin [10]. Most members of this family have 7 transmembrane helices.

A comparison between KXA homologue Brh1 (A0A091H593) and the MR homologue Bpr1 (K8F6P9) displayed an alignment between TMSs 2-4 of Brh1 and TMSs 6-8 of Bpr1, with a score of 16 S.D. (Figure 15). The full sequence comparison of Brh1 with the KXA protein, TC# 2.A.112.1.14 gave a score of 150 S.D. while the full sequence alignment of Bpr1 with MR protein, TC# 3.E.1.1.5 resulted in a score of 18 S.D. Based on the score of 16 S.D., the KXA family must be homologous to the MR family and other members of the TOG superfamily.

The KXA family (TC# 2.A.112) and the Nicotinamide Ribonucleoside (NR) Uptake Permease (PnuC) family (TC# 4.B.1)

The PnuC family is composed of membrane proteins found exclusively in bacteria with 7 predicted TMSs on average. The members of this family are known to play a role in transcriptional regulation and NR phosphorylation [5, 19, 20].

A comparison between the KXA homologue Tca2 (A0A0B2UUP4) and the PnuC homologue Pma1 (A0A090Y837) displayed an alignment of TMSs 1-3 of Tca2 with TMSs 2-5 of Pma1with a score of 14 S.D. (Figure 16). Interestingly, the first TMS of Tca1 aligned with TMSs 2-3 in Pma1. In addition, a score of 49 S.D. was obtained from the full sequence alignment of Tca2 and KXA protein, TC# 2.A.112.1.10 (E9FYJ5), while a score of 21 S.D. was obtained from the full length alignment of Pma1 and PnuC protein, TC# 4.B.1.1.6. These results suggest but do not establish that the KXA and the PnuC families are homologous.

The KXA family (TC# 2.A.112) and the G-protein-coupled receptor (GPCR) Family (TC# 9.A.14)

The GPCR family is involved in different types of signal transduction pathways which can be triggered by ligands such as hormones, peptides, odorants, and other subtances. Despite the high amount of sequence diversity within this family, its members have on average 7 TMSs.

A comparison between the KXA homologue Cmy1(XP_007061801.1) and the GPCR homologue Cfl3 (XP_014217479.1) displayed an alignment of their last 4 TMSs with a score of 15 S.D. (Figure 17). TMSs 4-7 of Cmy1 and TMSs 3-6 of Cfl3 aligned. In addition, a score of 37 S.D. was obtained when comparing the full sequences of Cmy1 with KXA protein, TC# 2.A.112.1.3, while a full sequence comparison of Cfl3 and GPCR protein, TC# 9.A.14.15.2 yielded a score of 88 S.D. These findings provide strong evidence that the KXA family is related to the GPCR family, and serve as yet another example where the N terminal TMS of KXA family members align with N terminal TMS of members of the TOG superfamily, further strengthening the argument that the KXA family is a member of the TOG superfamily.

Chapter 3: Other proteins homologous to the KX Blood-group Antigen (KXA) Family

3.1: Homology between the KX Blood-group Antigen (KXA) family and the Putative Heme Handling Protein (HHP) Family

The KXA family (TC# 2.A.112) and the Putative Heme Handling Protein (HHP) Family (TC# 9.B.14)

Members of the HHP family have a wide variety in their number of predicted TMSs, and are found in bacteria, archaea, and eukaryotes. Although the function of all these family members has not yet been established, it has been theorized that their main function may involve the transport of heme [2].

A comparison between the KXA homologue Cgr1 (XP_003505569.1) and the HHP homologue Ran1 (WP_004919909.1) showed an alignment of the last 4 TMSs for each protein (TMSs 6-9 in both proteins) with a score of 15 S.D. (Figure 18). In addition, a full sequence alignment between Cgr1 and the KXA protein, TC# 2.A.112.1.17 yielded a score of 54 S.D. while a full sequence alignment between Ran1 and HHP protein, TC# 9.B.14.3.2 gave a score of 43 S.D. these results satisfy our statistical requirements for homology and thus, provide strong evidence for the homologous relationship between the KX Blood-group Antigen (KXA) family and the Putative Heme Handling Protein (HHP) Family. These observations suggest that the HHP family is a member of the TOG superfamily.

Discussion

This study has provided evidence for the existence of 4 distinct subfamilies within the Kell Blood-group Antigen (KXA) family (Figures 1-6). Average hydropathy, amphipathicity, and similarity plots (Figure 8) suggested 9 well conserved TMSs in all or most of these subfamilies. KXA family members have been identified not only in metazoans, but also in viridiplantae, and intramacronucleata. The phylogenetic analysis of the 4 distinct subfamilies within the large KXA family revealed 17 primary clusters. Clusters 1-11 and 15-17 are formed by proteins belonging to subfamily 1, Cluster 13 is formed by proteins belonging to family 2, Cluster 14 is formed by proteins belonging to subfamily 3, and Cluster 12 is formed by proteins that belong to subfamily 4.

This study also established homology between the KXA family and 7 members of the Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily. This data strongly suggests that the KXA family is an additional member of this superfamily. The largest comparison scores were found with members of the sugar-transporting Sweet family. This relationship may serve as an indicator of the primary transport function of XK-related proteins. Interestingly, the last 4 TMSs in XK-related proteins could be shown to be homologous to the 4 TMS repeat unit in TOG superfamily proteins. This evidence led to the suggestion that a duplication of the 4 TMS unit had occurred during the evolution of these proteins. This fundamental suggestion will be examined in further studies involving the XK evolutionary pathway.

Lastly, homology was also found between the KXA family and the Putative Heme Handling Protein (HHP) Family, with the greatest similarities occurring in the last 4 TMSs of both proteins. These findings further suggest that the HHP family might also be a member of the TOG superfamily.

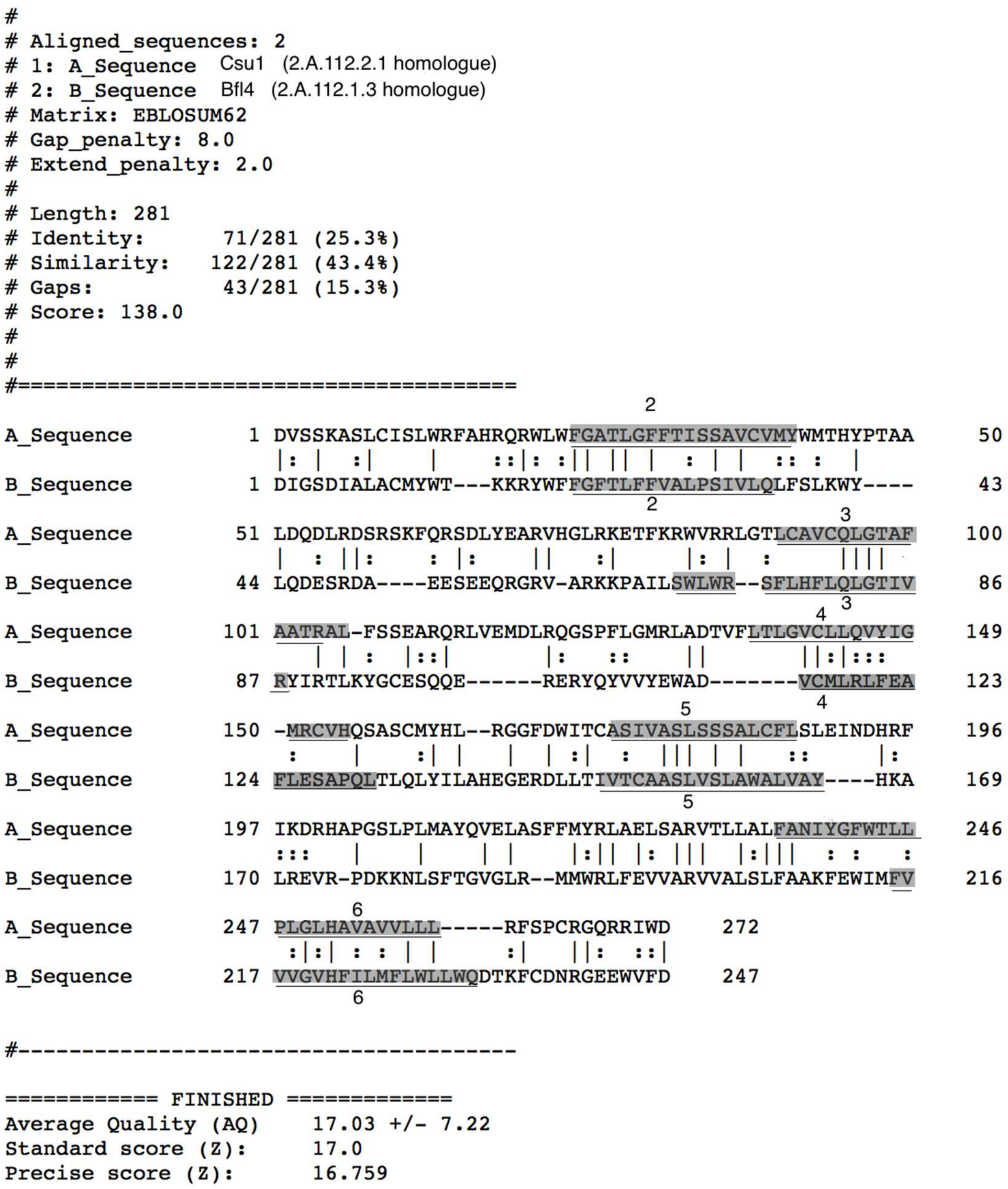
Table 1: Proving homology between the KX Blood-group Antigen (KXA) subfamilies.

Subfamilies compared	Protein A	Protein B	Protein C	Protein D	A v B	B v C	C v D	A v D
Sub1 v Sub2	Q8C0T0	C3ZRW5	I0Z719	A4RVV9	46	17	94	1
Sub1 v Sub3	Q8C0T0	XP_013411986.1	G0QX98	I7MG83	28	13	49	1
Sub1 v Sub4	Q8C0T0	A0A067R5K1	XP_012548325.1	B4J9J5	27	21	63	0.8
Sub2 v Sub3	K8EHG2	W4Y075	I7MG83	G0QX98	9*	16	46	5
Sub2 v Sub4	A4RVV9	A0A087UTH8	XP_008544814.1	B4J9J5	6*	21	61	0.4
Sub3 v Sub4	I7MG83	I7MG83	XP_013133483.1	B4J9J5	266	15	70	3

* Further analyses presented in Table 2 indicated that these proteins are homologous.

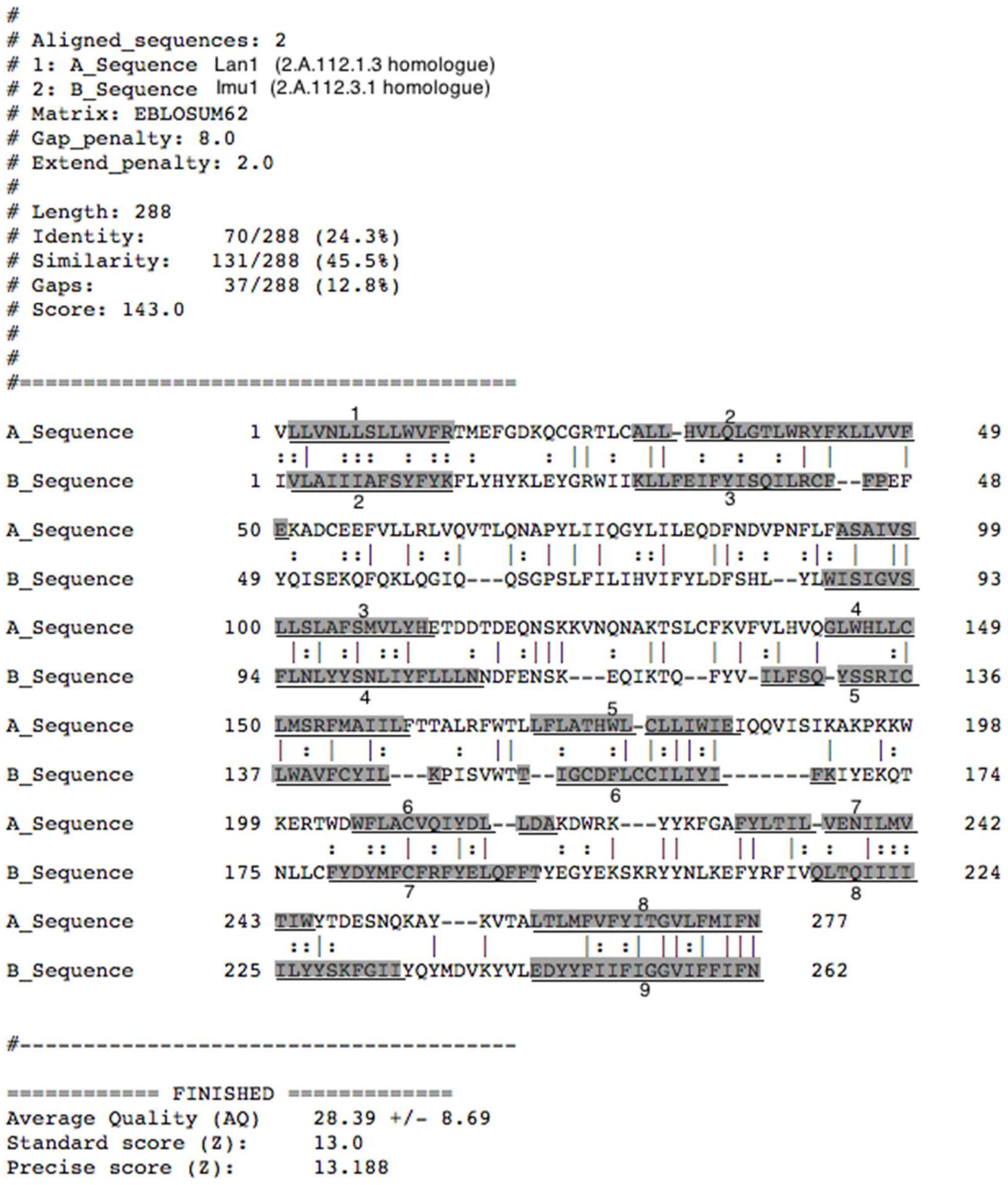
Table 2: Establishing homology between members of Table 1.

Protein A	Protein B	Protein C	Protein D	A v B	B v C	C v D	A v D
K8EHG2	I0Z719	XP_008426645.1	W4Y075	148	16	126	9
A4RVV9	I0Z719	C3ZRW5	A0A087UTH8	115	17	21	6



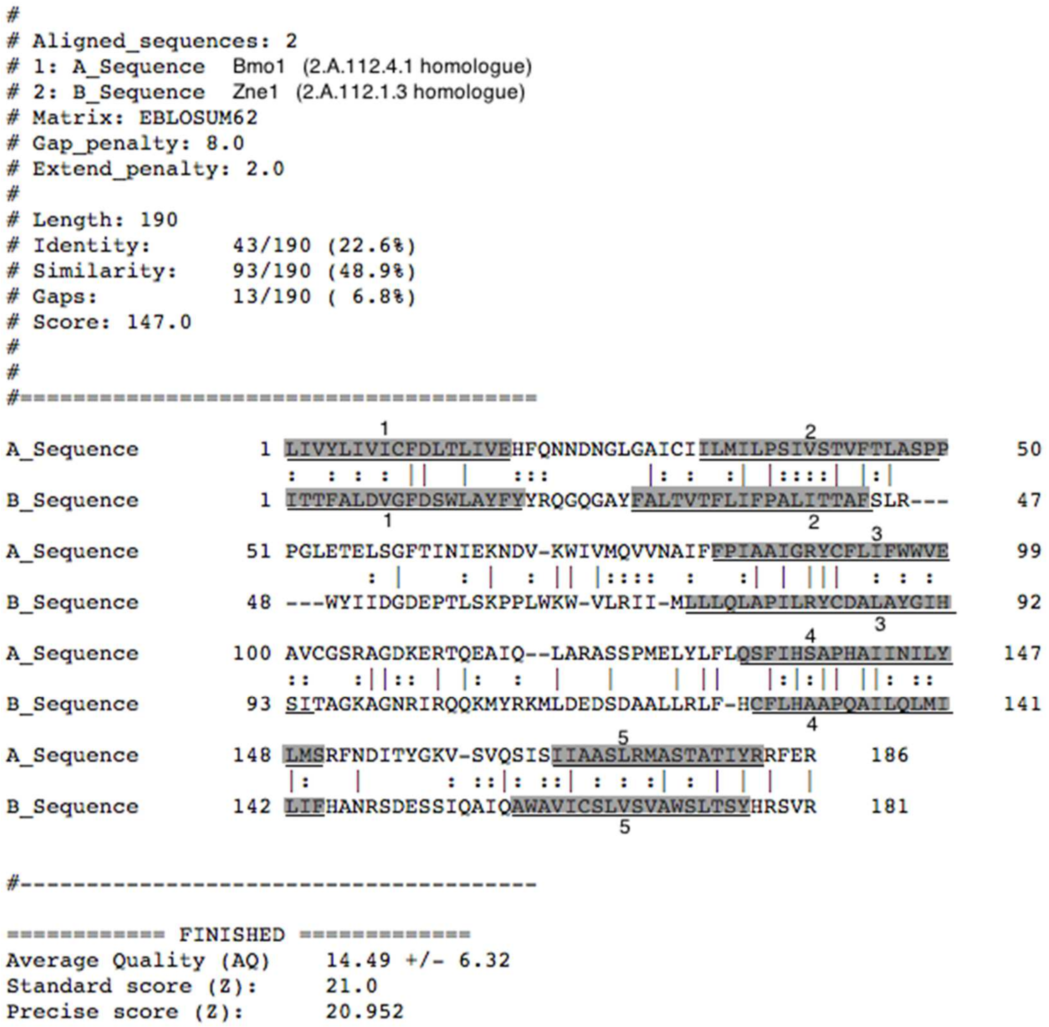
```
#
# Aligned_sequences: 2
# 1: A_Sequence Csu1 (2.A.112.2.1 homologue)
# 2: B_Sequence Bfl4 (2.A.112.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 281
# Identity:      71/281 (25.3%)
# Similarity:   122/281 (43.4%)
# Gaps:          43/281 (15.3%)
# Score: 138.0
#
#
#=======================================
                                                  2
A_Sequence         1 DVSSKASLCISLWRFAHRQRWLWFGATLGFFTISSAVCVMYWMTHYPTAA      50
                     |: |  :|    |    ::|: :|| || |  : | |  :: :  |
B_Sequence         1 DIGSDIALACMYWT---KKRYWFFGFTLFFVALPSIVLQLFSLKWY----      43
                                                  2
                                                                 3
A_Sequence        51 LDQDLRDSRSKFQRSDLYEARVHGLRKETFKRWVRRLGTLCAVCQLGTAF     100
                     |  : ||:    : |:    ||   :|     |: |  :     ||||
B_Sequence        44 LQDESRDA----EESEEQRGRV-ARKKPAILSWLWR--SFLHFLQLGTIV      86
                                                                 3
                                                               4
A_Sequence       101 AATRAL-FSSEARQRLVEMDLRQGSPFLGMRLADTVFLTLGVCLLQVYIG     149
                        | | :  |::|       |:   ::    ||       ||:|:::
B_Sequence        87 RYIRTLKYGCESQQE------RERYQYVVYEWAD-------VCMLRLFEA     123
                                                               4
                                                    5
A_Sequence       150 -MRCVHQSASCMYHL--RGGFDWITCASIVASLSSSALCFLSLEINDHRF     196
                      :    |    :| |   |  | :|  :  ||| | |   ::     |:
B_Sequence       124 FLESAPQLTLQLYILAHEGERDLLTIVTCAASLVSLAWALVAY----HKA     169
                                                     5
A_Sequence       197 IKDRHAPGSLPLMAYQVELASFFMYRLAELSARVTLLALFANIYGFWTLL     246
                     :::   |    |    | |    |:|| |: |||  |:|||  : :   :
B_Sequence       170 LREVR-PDKKNLSFTGVGLR--MMWRLFEVVARVVALSLFAAKFEWIMFV     216
                           6
A_Sequence       247 PLGLHAVAVVLLL-----RFSPCRGQRRIWD     272
                      :|:| : : | |     :|   ||:  ::|
B_Sequence       217 VVGVHFILMFLWLLWQDTKFCDNRGEEWVFD     247
                           6

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     17.03 +/- 7.22
Standard score (Z):      17.0
Precise score (Z):       16.759
```

Figure 1: GSAT alignment comparing subfamily 1 homologue, Bfl4, and subfamily 2 homologue, Csu1.



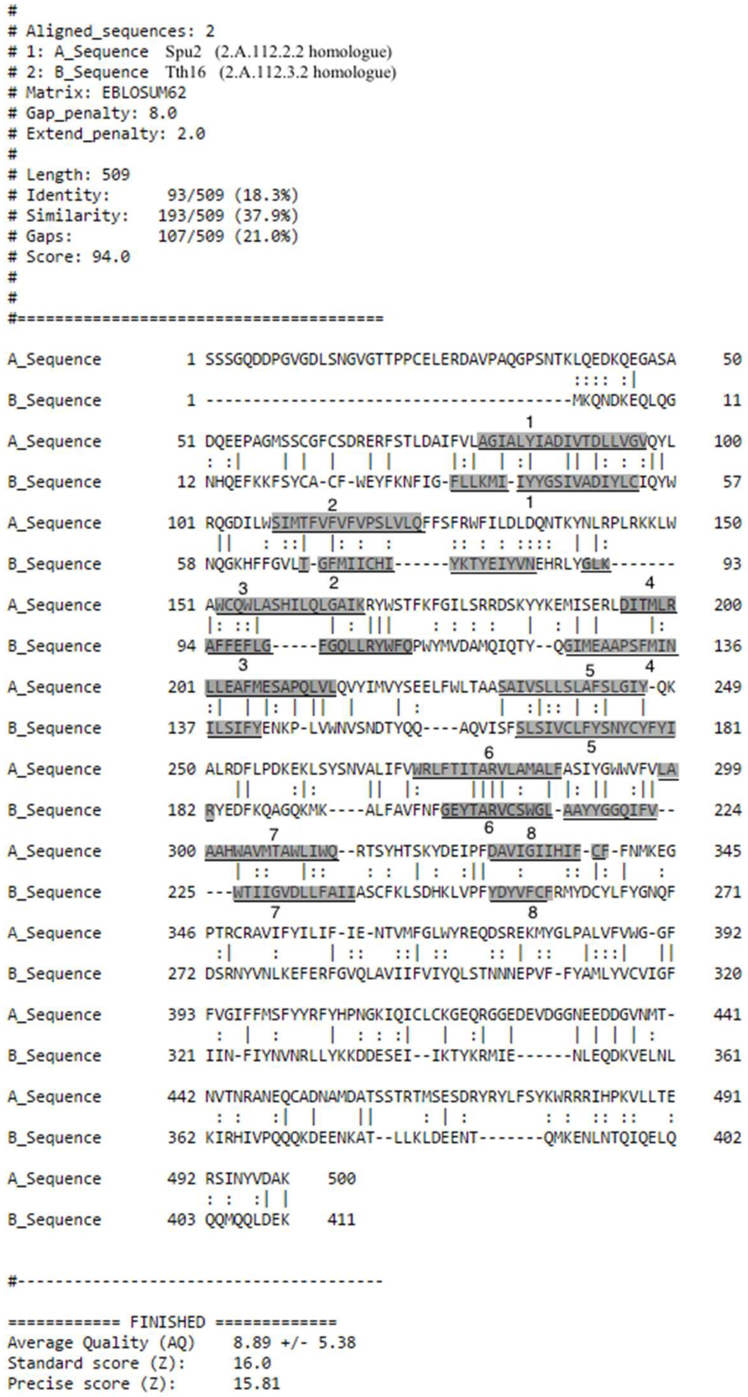
```
#
# Aligned_sequences: 2
# 1: A_Sequence Lan1  (2.A.112.1.3 homologue)
# 2: B_Sequence Imu1  (2.A.112.3.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 288
# Identity:      70/288 (24.3%)
# Similarity:   131/288 (45.5%)
# Gaps:          37/288 (12.8%)
# Score: 143.0
#
#
#=======================================

                          1                            2
A_Sequence        1 VLLVNLLSLLWVFRTMEFGDKQCGRTLCALL-HVLQLGTLWRYFKLLVVF      49
                    ::|  :::  : :: :     : || :  ||  :  :  : | |     |
B_Sequence        1 IVLAIIIAFSYFYKFLYHYKLEYGRWIIKLLFEIFYISQILRCF--FPEF      48
                          2                            3

A_Sequence       50 EKADCEEFVLLRLVQVTLQNAPYLIIQGYLILEQDFNDVPNFLFASAIVS      99
                     :   ::|  |: :|   |: | | |  ::|   ||: :  :|: |  ||
B_Sequence       49 YQISEKQFQKLQGIQ---QSGPSLFILIHVIFYLDFSHL--YLWISIGVS      93

                          3                                      4
A_Sequence      100 LLSLAFSMVLYHETDDTDEQNSKKVNQNAKTSLCFKVFVLHVQGLWHLLC     149
                     |:| :| ::|    : | :|||   :  ||   | | :|  |     :|
B_Sequence       94 FLNLYYSNLIYFLLLNNDFENSK---EQIKTQ--FYV-ILFSQ-YSSRIC     136
                          4                                      5
                                              5
A_Sequence      150 LMSRFMAIILFTTALRFWTLLFLATHWL-CLLIWIEIQQVISIKAKPKKW     198
                    | : |  |:     :  ||   :   :| |:||:|        |   |:
B_Sequence      137 LWAVFCYIL---KPISVWTT--IGCDFLCCILIYI-------FKIYEKQT     174
                                              6
                              6                                  7
A_Sequence      199 KERTWDWFLACVQIYDL--LDAKDWRK---YYKFGAFYLTIL-VENILMV     242
                        :  :: | : |:|     : : |   ||    ||  |: :  |:::
B_Sequence      175 NLLCFYDYMFCFRFYELQFFTYEGYEKSKRYYNLKEFYRFIVQLTQIIII     224
                              7                                  8
                                                8
A_Sequence      243 TIWYTDESNQKAY---KVTALTLMFVFYITGVLFMIFN    277
                     ::|:       |   |       |: :| ||:| |||
B_Sequence      225 ILYYSKFGIIYQYMDVKYVLEDYYFIIFIGGVIFFIFN    262
                                                9

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     28.39 +/- 8.69
Standard score (Z):      13.0
Precise score (Z):       13.188
```

Figure 2: GSAT alignment comparing subfamily 1 homologue, Lan1, and subfamily 3 homologue, Imu1.



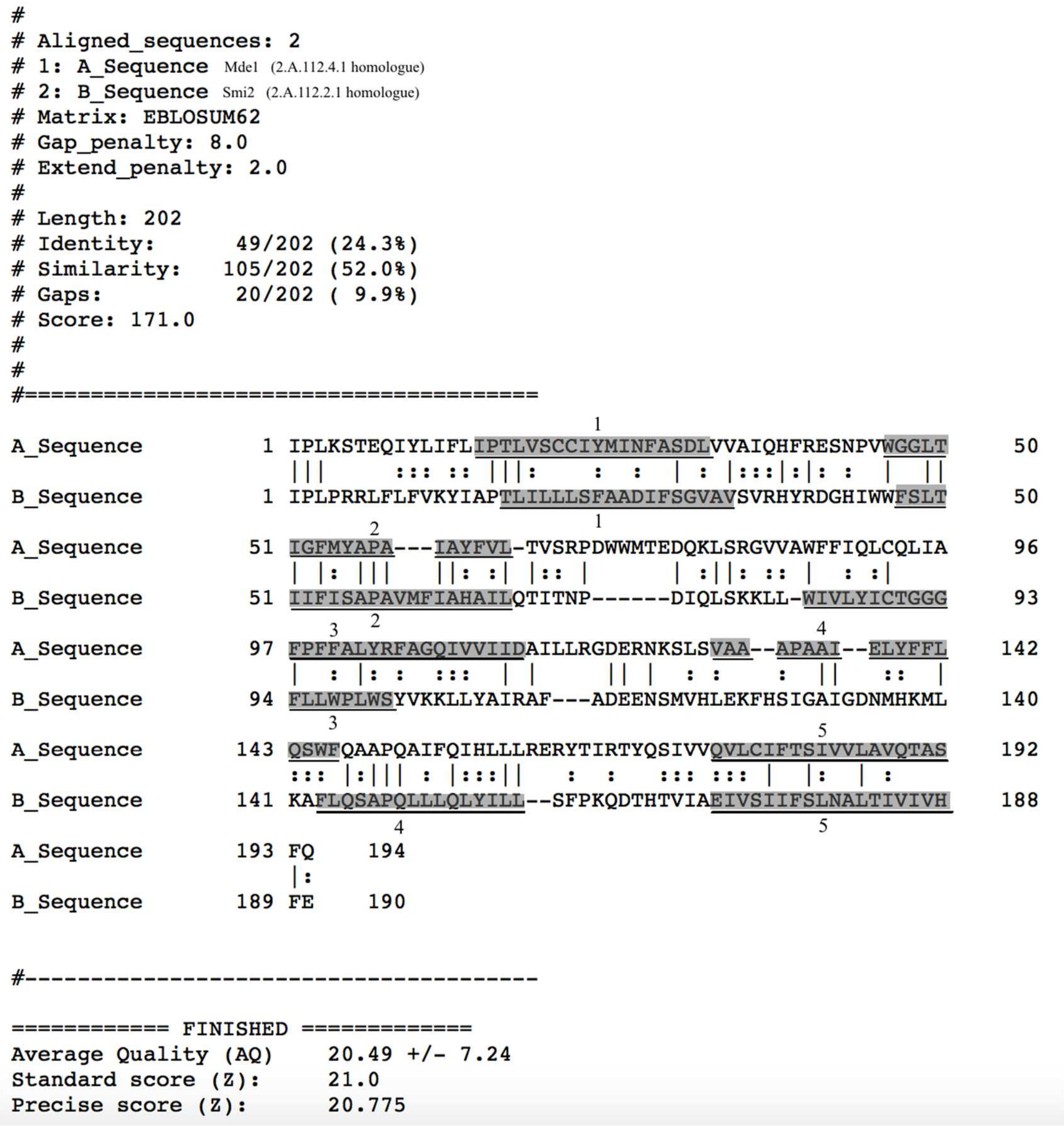
```
#
# Aligned_sequences: 2
# 1: A_Sequence   Bmo1  (2.A.112.4.1 homologue)
# 2: B_Sequence   Zne1  (2.A.112.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 190
# Identity:      43/190 (22.6%)
# Similarity:    93/190 (48.9%)
# Gaps:          13/190 ( 6.8%)
# Score: 147.0
#
#
#=======================================

                            1                               2
A_Sequence         1 LIVYLIVICFDLTLIVEHFQNNDNGLGAICIILMILPSIVSTVFTLASPP     50
                     :  : : : ||  |   :::       |: :  :| |::::| |:|
B_Sequence         1 ITTFALDVGFDSWLAYFYYRQGQGAYFALTVTFLIFPALITTAFSLR---     47
                            1                               2
                                                                 3
A_Sequence        51 PGLETELSGFTINIEKNDV-KWIVMQVVNAIFFPIAAIGRYCFLIFWWVE     99
                           : |    : |  : || |::::  :   :| | |||  : : :
B_Sequence        48 ---WYIIDGDEPTLSKPPLWKW-VLRII-MLLLQLAPILRYCDALAYGIH     92
                                                                 3
                                                            4
A_Sequence       100 AVCGSRAGDKERTQEAIQ--LARASSPMELYLFLQSFIHSAPHAIINILY    147
                     ::   :||:: | |:  :  |   |    | ||   |:|:|| ||: ::
B_Sequence        93 SITAGKAGNRIRQQKMYRKMLDEDSDAALLRLF-HCFLHAAPQAILQLMI    141
                                                            4
                                              5
A_Sequence       148 LMSRFNDITYGKV-SVQSISIIAASLRMASTATIYRRFER    186
                     |:   |      : ::|: ::| : : :| : | | |  |
B_Sequence       142 LIFHANRSDESSIQAIQAWAVICSLVSVAWSLTSYHRSVR    181
                                              5

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     14.49 +/- 6.32
Standard score (Z):      21.0
Precise score (Z):       20.952
```

Figure 3: GSAT Alignment comparing subfamily 1 homologue, Zne1, and subfamily 3 homologue, Bmo1.



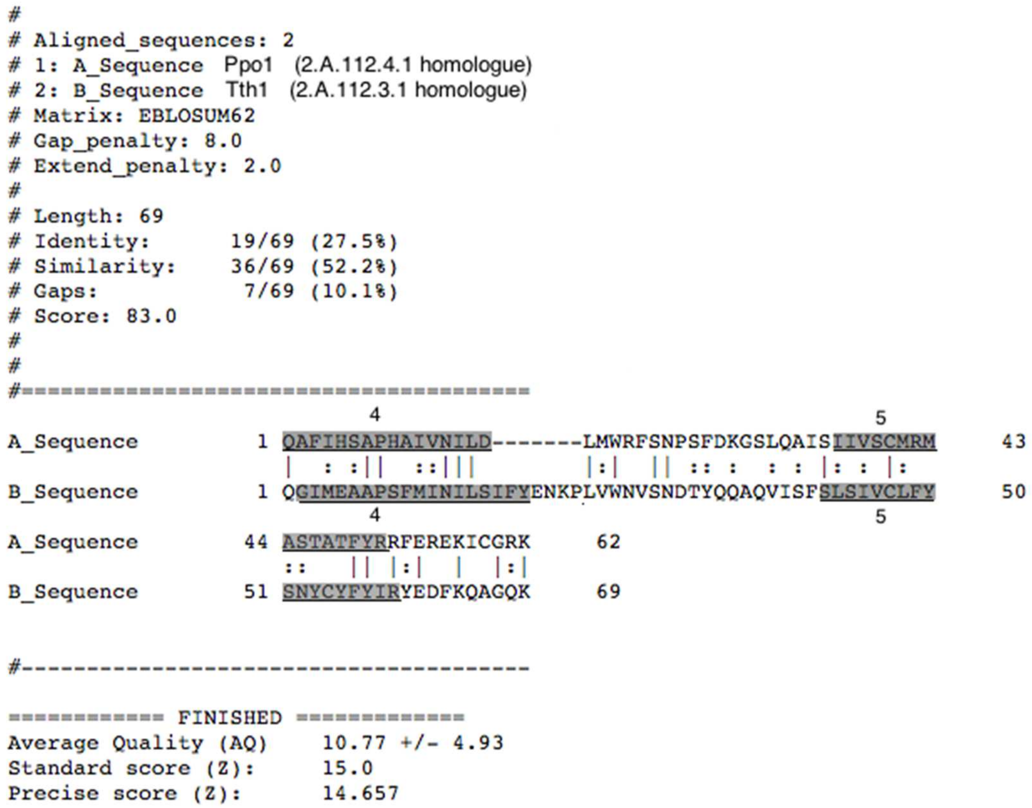
```
#
# Aligned_sequences: 2
# 1: A_Sequence  Spu2  (2.A.112.2.2 homologue)
# 2: B_Sequence  Tth16  (2.A.112.3.2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 509
# Identity:      93/509 (18.3%)
# Similarity:   193/509 (37.9%)
# Gaps:         107/509 (21.0%)
# Score: 94.0
#
#
#=======================================

A_Sequence         1 SSSGQDDPGVGDLSNGVGTTPPCELERDAVPAQGPSNTKLQEDKQEGASA     50
                                                            :::: :|
B_Sequence         1 ---------------------------------------MKQNDKEQLQG     11
                                                       1
A_Sequence        51 DQEEPAGMSSCGFCSDRERFSTLDAIFVLAGIALYIADIVTDLLVGVQYL    100
                     : :|    | |  |   | |      |:|  | :|   || |: : :||
B_Sequence        12 NHQEFKKFSYCA-CF-WEYFKNFIG-FLLKMI-IYYGSIVADIYLCIQYW     57
                                  2                    1
A_Sequence       101 RQGDILWSIMTFVFVFVPSLVLQFFSFRWFILDLDQNTKYNLRPLRKKLW    150
                      ||   : ::|  |: :  :      :: : : ::::  | |:
B_Sequence        58 NQGKHFFGVLT-GFMIICHI------YKTYEIYVNEHRLYGLK-------     93
                        3         2                                 4
A_Sequence       151 AWCQWLASHILQLGAIKRYWSTFKFGILSRRDSKYYKEMISERLDITMLR    200
                     |: ::|       | : |||    : : : :   |  : | |     |:
B_Sequence        94 AFFEFLG-----FGQLLRYWFQPWYMVDAMQIQTY--QGIMEAAPSFMIN    136
                        3                                    5      4
A_Sequence       201 LLEAFMESAPQLVLQVYIMVYSEELFWLTAASAIVSLLSLAFSLGIY-QK    249
                     :|  | |: | ||  |    | :        |  :|:: | :|   |
B_Sequence       137 ILSIFYENKP-LVWNVSNDTYQQ----AQVISFSLSIVCLFYSNYCYFYI    181
                                                   6         5
A_Sequence       250 ALRDFLPDKEKLSYSNVALIFVWRLFTITARVLAMALFASIYGWVFVLA    299
                        ||    :|:     ||  |:     |||| :  |  |: ||  :||
B_Sequence       182 RYEDFKQAGQKMK----ALFAVFNFGEYTARVCSWGL-AAYYGGQIFV--    224
                            7                      6   8
A_Sequence       300 AAHWAVMTAWLIWQ--RTSYHTSKYDEIPFDAVIGIIHIF-CF-FNMKEG    345
                        | ::   |::    : :  | :  :||   :    :: |: |   :
B_Sequence       225 ---WTIIGVDLLFAIIASCFKLSDHKLVPFYDYVFCFRMYDCYLFYGNQF    271
                            7                          8
A_Sequence       346 PTRCRAVIFYILIF-IE-NTVMFGLWYREQDSREKMYGLPALVFVWG-GF    392
                      :|    :     | ::   ::| ::    :: | ::   |:::|   ||
B_Sequence       272 DSRNYVNLKEFERFGVQLAVIIFVIYQLSTNNNEPVF-FYAMLYVCVIGF    320

A_Sequence       393 FVGIFFMSFYYRFYHPNGKIQICLCKGEQRGGEDEVDGGNEEDDGVNMT-    441
                      :  |  :     |  : : :|   |  :|  |      | | | | :
B_Sequence       321 IIN-FIYNVNRLLYKKDDESEI--IKTYKRMIE------NLEQDKVELNL    361

A_Sequence       442 NVTNRANEQCADNAMDATSSTRTMSESDRYRYLFSYKWRRRIHPKVLLTE    491
                      : :   :|  |    ||     : | :        : :  :: ::   :
B_Sequence       362 KIRHIVPQQQKDEENKAT--LLKLDEENT-------QMKENLNTQIQELQ    402

A_Sequence       492 RSINYVDAK    500
                     : :  :| |
B_Sequence       403 QQMQQLDEK    411


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)    8.89 +/- 5.38
Standard score (Z):     16.0
Precise score (Z):      15.81
```

Figure 4: GSAT Alignment comparing subfamily 2 homologue, Spu2, and subfamily 3 homologue, Tth16.



```
#
# Aligned_sequences: 2
# 1: A_Sequence Mde1 (2.A.112.4.1 homologue)
# 2: B_Sequence Smi2 (2.A.112.2.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 202
# Identity:      49/202 (24.3%)
# Similarity:   105/202 (52.0%)
# Gaps:          20/202 ( 9.9%)
# Score: 171.0
#
#
#=======================================

A_Sequence         1 IPLKSTEQIYLIFLIPTLVSCCIYMINFASDLVVAIQHFRESNPVWGGLT     50
                     |||     ::: ::  |||:    :  :  | : |:::|:|: :  |  ||
B_Sequence         1 IPLPRRLFLFVKYIAPTLILLLSFAADIFSGVAVSVRHYRDGHIWWFSLT     50

A_Sequence        51 IGFMYAPA---IAYFVL-TVSRPDWWMTEDQKLSRGVVAWFFIQLCQLIA     96
                     | |: |||   ||: :| |:: |      | :||: :: |  : :|
B_Sequence        51 IIFISAPAVMFIAHAILQTITNP------DIQLSKKLL-WIVLYICTGGG     93

A_Sequence        97 FPFFALYRFAGQIVVIIDAILLRGDERNKSLSVAA--APAAI--ELYFFL    142
                     |  : |: :  :::  | |     || |  : :    :  ||   ::  |
B_Sequence        94 FLLWPLWSYVKKLLYAIRAF---ADEENSMVHLEKFHSIGAIGDNMHKML    140

A_Sequence       143 QSWFQAAPQAIFQIHLLLRERYTIRTYQSIVVQVLCIFTSIVVLAVQTAS    192
                     ::: |:||| : |:::||   :  :   ::: ::: |  |:  | :
B_Sequence       141 KAFLQSAPQLLLQLYILL--SFPKQDTHTVIAEIVSIIFSLNALTIVIVH    188

A_Sequence       193 FQ    194
                     |:
B_Sequence       189 FE    190


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)    20.49 +/- 7.24
Standard score (Z):     21.0
Precise score (Z):      20.775
```

Figure 5: GSAT alignment comparing subfamily 2 homologue, Smi2, and subfamily 4 homologue, Mde1.



```
#
# Aligned_sequences: 2
# 1: A_Sequence Ppo1  (2.A.112.4.1 homologue)
# 2: B_Sequence Tth1  (2.A.112.3.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 69
# Identity:      19/69 (27.5%)
# Similarity:    36/69 (52.2%)
# Gaps:           7/69 (10.1%)
# Score: 83.0
#
#
#=======================================
                            4                                     5
A_Sequence         1 QAFIHSAPHAIVNILD-------LMWRFSNPSFDKGSLQAISIIVSCMRM     43
                     |  : :||  ::|||        |:|  || :: :  : : |: : |:
B_Sequence         1 QGIMEAAPSFMINILSIFYENKPLVWNVSNDTYQQAQVISFSLSIVCLFY     50
                            4                                     5
A_Sequence        44 ASTATFYRRFEREKICGRK     62
                     ::   || |:|  |  |:|
B_Sequence        51 SNYCYFYIRYEDFKQAGQK     69


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)    10.77 +/- 4.93
Standard score (Z):     15.0
Precise score (Z):      14.657
```

Figure 6: GSAT alignment comparing subfamily 3 homologue, Tth1, and subfamily 4 homologue, Ppo1.

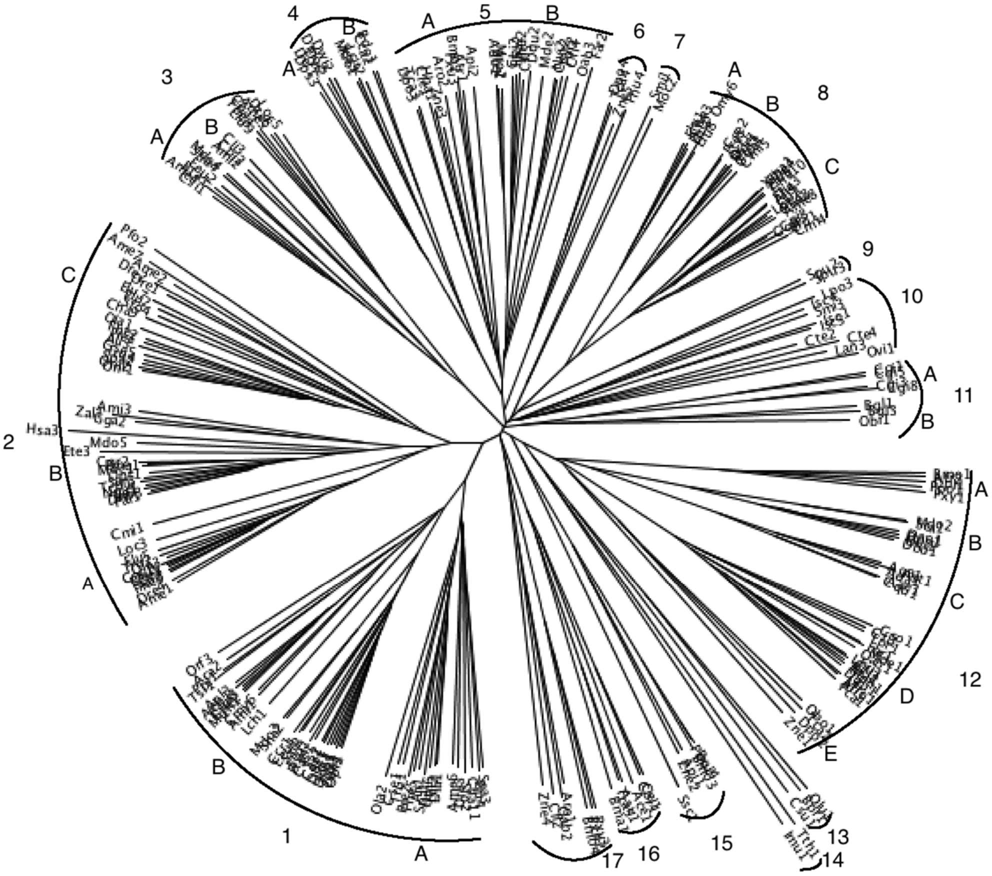


Figure 7: Phylogenetic tree of the 248 KX Blood-group Antigen (KXA) family proteins included in this study.

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree.

Cluster 1A			Average: 378 +/- 20			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	Organism name	Common name
1	Ssa3	929277123	349	7	*Salmo salar*	Atlantic salmon
2	Elu4	742165662	342	9	*Esox lucius*	Northern pike
3	Cha11	831280934	354	6	*Clupea harengus*	Atlantic herring
4	Sfo2	938076571	361	8	*Scleropages formosus*	Asian arowana
5	Dre3	92097675	378	9	*Danio rerio*	Zebrafish
6	Ame6	597797192	379	7	*Astyanax mexicanus*	Mexican tetra
7	Dla1	317419881	382	8	*Dicentrarchus labrax*	European seabass
8	Lcr2	734595575	427	9	*Larimichthys crocea*	Yellow croaker
9	Oni6	542193682	386	8	*Oreochromis niloticus*	Nile tilapia
10	Spa1	657574274	393	9	*Stegastes partitus*	Bicolor damselfish
11	Ali1	928019446	375	9	*Austrofundulus limnaeus*	Ray-finned fish
12	Fhe5	831507323	381	8	*Fundulus heteroclitus*	Killifish
13	Pre2	658908969	400	8	*Poecilia reticulata*	Rainbow fish

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

14	Tru1	74096405	381	9	*Takifugu rubripes*	Pufferfish
15	Cse1	657741710	379	7	*Cynoglossus semilaevis*	Tongue-sole
16	Ola2	432930062	379	6	*Oryzias latipes*	Japanese rice fish
Cluster 1B			Average: 377 +/- 7			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
17	Opr2	504151511	373	8	*Ochotona princeps*	American pika
18	Ode1	507669953	373	9	*Octodon degus*	Degu
19	Tsy1	640812260	373	8	*Tarsius syrichta*	Philippine tarsier
20	Dno1	488515378	392	9	*Dasypus novemcinctus*	Nine-banded armadillo
21	Laf1	344273150	373	9	*Loxodonta africana*	African bush elephant
22	Dor1	852784758	372	9	*Dipodomys ordii*	Ord's kangaroo rat
23	Cgr1	354486806	373	9	*Cricetulus griseus*	Chinese hamster
24	Jja2	847003305	373	9	*Jaculus jaculus*	Egyptian jerboa
25	Cas1	586458743	372	9	*Chrysochloris asiatica*	Cape golden mole
26	Ete4	507624678	372	9	*Echinops telfairi*	Lesser hedgehog tenrec
27	Eed1	585640774	373	8	*Elephantulus edwardii*	Cape elephant shrew
28	Sar2	505776918	371	9	*Sorex araneus*	Commonshrew

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

29	Eeu1	617550849	373	7	*Erinaceus europaeus*	European hedgehog
30	Ccr1	507938340	367	8	*Condylura cristata*	Star-nosed mole
31	Sha2	82146472	390	8	*Sarcophilus harrisii*	Tasmanian devil
32	Mdo3	612005370	392	9	*Monodelphis domestica*	Gray short-tailed opossum
33	Lch1	556967510	375	9	*Latimeria chalumnae*	West Indian Ocean coelacanth
34	Ami1	564234535	377	9	*Alligator mississippiensis*	American alligator
35	Cmy6	591358158	378	9	*Chelonia mydas*	Green sea turtle
36	Mun1	704582202	380	10	*Mesitornis unicolor*	Brown mesite
37	Cpe2	701385195	382	9	*Chaetura pelagica*	chimney swift
38	Can3	663290624	377	9	*Calypte anna*	Anna's hummingbird
39	Ach1	677602240	380	7	*Acanthisitta chloris*	Rifleman
40	Apl3	514753082	378	9	*Anas platyrhynchos*	Wild duck
41	Tsi1	927148170	371	9	*Thamnophis sirtalis*	Common garter snake
42	Pbi1	602665306	388	9	*Python bivittatus*	Burmese python
43	Aca2	327269813	379	9	*Anolis carolinensis*	Carolina anole
44	Orf3	76780849	367	9	*Xenopus (Silurana) tropicalis*	Western clawed frog

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

Cluster 2A			Average: 414 +/- 10			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	Organism	Common name
45	Ame1	597782502	408	8	*Astyanax mexicanus*	Blind cave fish
46	Dre4	237820640	417	9	*Danio rerio*	Zebrafish
47	Fhe9	831497743	412	9	*Fundulus heteroclitus*	Mummichog
48	Ali2	928061520	421	9	*Austrofundulus limnaeus*	Ray-finned fish
49	Cse3	657804438	421	10	*Cynoglossus semilaevis*	Three-lined tongue sole
50	Oni5	348519554	407	9	*Oreochromis niloticus*	Nile tilapia
51	Gac1	390407687	404	9	*Gasterosteus aculeatus*	Three-spined stickleback
52	Ola3	432882983	399	9	*Oryzias latipes*	Japanese rice fish
53	Tru2	768934133	421	9	*Takifugu rubripes*	Pufferfish
54	Elu5	742247765	406	9	*Esox lucius*	Northern pike
55	Loc3	573887401	429	9	*Lepisosteus oculatus*	Spotted gar
56	Cmi1	632954500	427	10	*Callorhinchus milii*	Australian ghostshark
Cluster 2B			Average: 406 +/- 17			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

57	Pal1	586539149	393	9	*Pteropus alecto*	Black fruit bat
58	Dno3	488527905	395	8	*Dasypus novemcinctus*	Nine-banded armadillo
59	Jja1	507575568	399	8	*Jaculus jaculus*	Lesser Egyptian jerboa
60	Nga1	674077531	384	9	*Nannospalax galili*	Blind mole rat
61	Tch1	947326929	399	8	*Tupaia chinensis*	Chinese tree shrew
62	Sar3	505781406	399	9	*Sorex araneus*	Common shrew
63	Opr1	504164626	397	8	*Ochotona princeps*	American pika
64	Moc1	532018865	399	9	*Microtus ochrogaster*	Prairie vole
65	Pma1	589925604	399	8	*Peromyscus maniculatus bairdii*	Deer mouse
66	Cgr2	354472347	400	9	*Cricetulus griseus*	Chinese hamster
67	Ete3	507633073	451	9	*Echinops telfairi*	Lesser hedgehog tenrec
68	Hsa3	767904886	413	5	*Homo sapiens*	Human
69	Mdo5	612024728	439	10	*Monodelphis domestica*	Gray short-tailed opossum
70	Gga2	113206170	414	7	*Gallus gallus*	Red junglefow
71	Zal1	542167755	406	7	*Zonotrichia albicollis*	White-throated sparrow
72	Ami3	944428637	407	10	*Alligator mississippiensis*	American alligator

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

Cluster 2C			Average: 416 +/- 32			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
73	Oni1	542227949	421	8	*Oreochromis niloticus*	Nile tilapia
74	Oni2	908527864	421	9	*Oreochromis niloticus*	Nile tilapia
75	Spa2	657560559	390	9	*Stegastes partitus*	Stegastes partitus
76	Tru4	410911598	413	9	*Takifugu rubripes*	pufferfish
77	Cse5	657804349	415	9	*Cynoglossus semilaevis*	Three-lined tongue sole
78	Ali3	928061523	434	9	*Austrofundulus limnaeus*	Ray-finned fish
79	Pre1	658876423	411	7	*Poecilia reticulata*	Rainbow fish
80	Fhe2	831497783	416	7	*Fundulus heteroclitus*	Mummichog
81	Ola1	765135066	413	9	*Oryzias latipes*	Japanese rice fish
82	Cha9	831309001	407	9	*Clupea harengus*	Atlantic herring
83	Sfo4	938082224	415	7	*Scleropages formosus*	Asian arowana
84	Elu2	884958195	402	8	*Esox lucius*	Northern pike
85	Elu7	742247767	409	10	*Esox lucius*	Northern pike
86	Dre1	528509979	396	8	*Danio rerio*	Zebrafish
87	Dre2	528509977	397	9	*Danio rerio*	Zebrafish

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

88	Ame2	597782504	390	9	*Astyanax mexicanus*	Mexican tetra
89	Ame7	597782543	409	10	*Astyanax mexicanus*	Mexican tetra
90	Pfo2	617481736	535	11	*Poecilia formosa*	Amazon molly
Cluster 3A			Average: 690 +/- 28			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
91	Csi1	478532983	688	6	*Ceratotherium simum simum*	White rhinoceros
92	Ame4	301789153	675	6	*Ailuropoda melanoleuca*	giant panda
93	Bta1	741903496	662	6	*Bos taurus*	Cattle
94	Eeu2	617574374	668	5	*Erinaceus europaeus*	European hedgehog
95	Sha4	395509490	717	8	*Sarcophilus harrisii*	Tasmanian devil
96	Mdo4	126303977	731	7	*Monodelphis domestica*	Gray short-tailed opossum
Cluster 3B			Average: 668			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
97	Ami2	944464095	619	8	*Alligator mississippiensis*	American alligator
98	Cpi2	641795697	737	8	*Chrysemys picta bellii*	Painted turtle
99	Cli3	915535437	647	7	*Columba livia*	Rock dove

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

Cluster 3C			Average: 649 +/- 29			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
100	Elu9	742179754	683	9	*Esox lucius*	Northern pike
101	Fhe8	831535968	646	10	*Fundulus heteroclitus*	Mummichog
102	Cha1	831316706	686	10	*Clupea harengus*	Atlantic herring
103	Ame5	597797989	635	9	*Astyanax mexicanus*	Mexican tetra
104	Dre6	59276054	622	7	*Danio rerio*	Zebrafish
105	Loc5	573910558	620	11	*Lepisosteus oculatus*	Spotted gar
Cluster 4A			Average: 352 +/- 5			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
106	Dps3	198469731	347	9	*Drosophila pseudoobscura*	Fruit fly
107	Der1	194893875	352	9	*Drosophila erecta*	West African fruit fly
108	Dvi2	195393902	347	9	*Drosophila virilis*	Fruit fly
109	Dmo2	195134448	355	8	*Drosophila mojavensis*	Fruit fly
110	Dwi2	946585539	359	8	*Drosophila willistoni*	Fruit fly

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

Cluster 4B			Average: 432 +/- 23			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
111	Mdo1	557772520	434	9	*Musca domestica*	Housefly
112	Sca3	907638182	443	9	*Stomoxys calcitrans*	Stable fly
113	Lcu2	906465221	464	9	*Lucilia cuprina*	Australian sheep blowfly
114	Cca2	498980850	410	9	*Ceratitis capitata*	Mediterranean fruit fly
115	Bdo1	751794742	410	7	*Bactrocera dorsalis*	Oriental fruit fly
Cluster 5A			Average: 406 +/- 21			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
116	Dpo3	332376603	399	8	*Dendroctonus ponderosae*	Mountain pine beetle
117	Tca3	270009684	393	9	*Tribolium castaneum*	Red flour beetle
118	Cle1	939256875	423	8	*Cimex lectularius*	Bed bug
119	Hha1	939636702	403	9	*Halyomorpha halys*	Stink bug
120	Zne3	646714532	371	9	*Zootermopsis nevadensis*	Dampwood termite
121	Aro2	817085449	432	9	*Athalia rosae*	Turnip sawfly

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

122	Pxu3	910324560	397	8	*Papilio xuthus*	Asian swallowtail
123	Bmo2	512905832	442	9	*Bombyx mori*	Silkworm
124	Atr1	913332118	397	9	*Amyelois transitella*	Navel orangeworm
125	Api2	328712602	403	10	*Acyrthosiphon pisum*	Pea aphid
Cluster 5B			Average: 428 +/- 11			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
126	Ado1	572299643	418	8	*Apis dorsata*	Giant honey bee
127	Bte1	340712395	422	10	*Bombus terrestris*	Buff-tailed bumblebee
128	Mro2	383856841	420	10	*Megachile rotundata*	Alfalfa leafcutter bee
129	Cbi2	759053870	425	7	*Cerapachys biroi*	Clonal raider ant
130	Pba2	769863968	424	8	*Pogonomyrmex barbatus*	Harvester ant
131	Lhu2	815803674	423	8	*Linepithema humile*	Argentine ant
132	Cfl5	752868587	420	8	*Camponotus floridanus*	Florida carpenter ant
133	Dqu2	951533429	436	8	*Dinoponera quadriceps*	South American ant
134	Mde2	665789477	447	9	*Microplitis demolitor*	No Common name
135	Nvi2	644999810	429	9	*Nasonia vitripennis*	Jewel wasps

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

136	Cso2	766931685	423	10	*Ceratosolen solmsi marchali*	Fig wasp
137	Tpr2	936710859	425	9	*Trichogramma pretiosum*	Parasitic wasp
138	Cfl4	936577548	440	9	*Copidosoma floridanum*	Wasp, no common name
139	Oab3	817197359	416	9	*Orussus abietinus*	Sawflies
140	Far2	755952701	453	10	*Fopius arisanus*	Melon fly
Cluster 6			Average: 423 +/- 3			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
141	Dpo4	478257949	427	9	*Dendroctonus ponderosae*	Mountain pine beetle
142	Tca4	91086089	422	9	*Tribolium castaneum*	Red flour beetle
143	Zne5	646714531	424	9	*Zootermopsis nevadensis*	Dampwood termite
144	Phu4	242009427	419	10	*Pediculus humanus corporis*	Body louse
Cluster 7			Average: 509 +/- 76			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
145	Smi1	675376550	455	9	*Stegodyphus mimosarum*	Social spiders

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

146	Moc2	391347797	563	7	*Metaseiulus occidentalis*	Predatory mite
Cluster 8A			Average: 575 +/- 9	Number of predicted TMSs		
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids		*Organism*	Common name
147	Elu1	742154134	584	7	*Esox lucius*	Northern pike
148	Cha3	831292566	574	9	*Clupea harengus*	Atlantic herring
149	Fhe6	831548101	568	7	*Fundulus heteroclitus*	Mummichog
150	Omy6	642128038	565	6	*Oncorhynchus mykiss*	Rainbow trout
151	Elu8	742241109	584	7	*Esox lucius*	Northern pike
Cluster 8B			Average: 636 +/- 129			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
152	Cse2	657811285	614	10	*Cynoglossus semilaevis*	Three-lined tongue sole
153	Pre2	658908969	400	8	*Poecilia reticulata*	Rainbow fish
154	Orf1	73853788	617	10	*Xenopus (Silurana) tropicalis*	Western clawed frog
155	Oor1	466034569	653	10	*Orcinus orca*	Killer whale
156	Fca1	755809350	822	10	*Felis catus*	Domestic cat
157	Gga1	513172751	726	10	*Gallus gallus*	Chicken

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

158	Cmi5	632947791	619	10	*Callorhinchus milii*	Ghostshark
Cluster 8C			Average: 570 +/- 35			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
159	Xma1	551494869	536	10	*Xiphophorus maculatus*	Southern platyfish
160	Fhe4	831559042	566	8	*Fundulus heteroclitus*	Mummichog
161	Elu6	742184651	564	6	*Esox lucius*	Northern pike
162	Cha10	831289534	569	8	*Clupea harengus*	Atlantic herring
163	Fhe3	831480618	614	8	*Fundulus heteroclitus*	Mummichog
164	Ali4	928068659	594	8	*Austrofundulus limnaeus*	Ray-finned fish
165	Elu3	884961059	598	8	*Esox lucius*	Northern pike
166	Loc2	573903915	558	8	*Lepisosteus oculatus*	Spotted gar
167	Cha8	831299310	572	8	*Clupea harengus*	Atlantic herring
168	Ame8	597774265	570	8	*Astyanax mexicanus*	Mexican tetra
169	Cpi3	641773414	538	8	*Chrysemys picta bellii*	Painted turtle
170	Gga3	363741524	489	7	*Gallus gallus*	Red junglefowl
171	Efu1	641701676	630	6	*Eptesicus fuscus*	Big brown bat

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

172	Cmi4	632940021	587	5	*Callorhinchus milii*	Australian ghostshark
Cluster 9			Average: 551 +/- 1			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
173	Spu2	72155137	551	9	*Strongylocentrotus purpuratus*	Purple sea urchin
174	Spu3	780027696	550	9	*Strongylocentrotus purpuratus*	Purple sea urchin
Cluster 10			Average: 459 +/- 83			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
175	Lpo3	926623730	500	9	*Limulus polyphemus*	Atlantic horseshoe crab
176	Isc4	241750940	463	8	*Ixodes scapularis*	Deer tick
177	Smi3	675369640	390	8	*Stegodyphus mimosarum*	Social spiders
178	Isc1	241628149	348	9	*Ixodes scapularis*	Deer tick
179	Isc3	241750942	437	9	*Ixodes scapularis*	Deer tick
180	Cte2	443733912	364	8	*Capitella teleta*	Bristle worms
181	Cte4	443710372	597	7	*Capitella teleta*	Bristle worms

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

182	Lan3	919075193	488	10	*Lingula anatina*	Lingula
183	Ovi1	684413160	540	9	*Opisthorchis viverrini*	Southeast Asian liver fluke
Cluster 11A			Average: 488 +/- 68			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
184	Cgi1	762148405	457	10	*Crassostrea gigas*	Pacific oyster
185	Cgi5	762148409	453	9	*Crassostrea gigas*	Pacific oyster
186	Cgi3	762169010	451	7	*Crassostrea gigas*	Pacific oyster
187	Cgi8	762148424	590	7	*Crassostrea gigas*	Pacific oyster
Cluster 11B			Average: 548 +/- 55			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
188	Bgl1	908401653	519	8	*Biomphalaria glabrata*	Freshwater snail
189	Bgl3	908401660	612	8	*Biomphalaria glabrata*	Freshwater snail
190	Obi1	918308283	514	9	*Octopus bimaculoides*	California two-spot octopus
Cluster 12A			Average: 586 +/- 8			

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
191	Bmo1	827553251	583	9	*Bombyx mori*	Silkworm
192	Atr2	913303002	590	9	*Amyelois transitella*	Navel orangeworm
193	Dpl1	357620498	597	9	*Danaus plexippus*	Monarch butterfly
194	Ppo1	909557061	580	9	*Papilio polytes*	Common Mormon
195	Pxy1	768445686	578	9	*Plutella xylostella*	Diamondback moth
Cluster 12B			Average: 654 +/- 11			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
196	Mdo2	557753615	663	9	*Musca domestica*	Housefly
197	Sca1	907691150	668	8	*Stomoxys calcitrans*	Stable fly
198	Dan1	194755753	651	8	*Drosophila ananassae*	Fruit fly (Variant)
199	Dps1	198460665	651	8	*Drosophila pseudoobscura*	Fruit fly (Variant)
200	Dvi1	195383778	653	8	*Drosophila virilis*	Fruit fly (Variant)
201	Dbu1	924557052	635	8	*Drosophila busckii*	Fruit fly (Variant)
Cluster 12C			Average: 662 +/- 65			

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
202	Aga1	347967575	634	8	*Anopheles gambiae str. PEST*	Mosquito (Variant)
203	Ast1	374720901	777	8	*Anopheles stephensi*	Mosquito (Variant)
204	Ada1	568253437	641	8	*Anopheles darlingi*	Mosquito (Variant)
205	Aae1	157116956	615	9	*Aedes aegypti*	Yellow fever mosquito
206	Cqu1	170044884	642	9	*Culex quinquefasciatus*	Southern house mosquito
Cluster 12D			Average: 535 +/- 10			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
207	Cso1	766928864	524	9	*Ceratosolen solmsi marchali*	Wasp (Variant)
208	Cfl1	936578914	535	9	*Copidosoma floridanum*	Wasp (Variant)
209	Tpr1	936702373	509	9	*Trichogramma pretiosum*	Wasp (Variant)
210	Far1	755940126	525	9	*Fopius arisanus*	No Common name
211	Mde1	665794251	534	9	*Microplitis demolitor*	No Common name

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

212	Oab1	817221116	541	9	*Orussus abietinus*	Sawfly
213	Lhu1	815818514	540	9	*Linepithema humile*	Argentine ant
214	Cfl3	752887474	538	9	*Camponotus floridanus*	Florida carpenter ant
215	Pba1	769849258	547	9	*Pogonomyrmex barbatus*	Harvester ant
216	Mph1	826408267	550	9	*Monomorium pharaonis*	Pharaoh ant
217	Dqu1	951536602	540	9	*Dinoponera quadriceps*	South American ant
218	Mro1	805752510	536	9	*Megachile rotundata*	Alfalfa leafcutter bee
219	Afl1	820859990	530	9	*Apis florea*	Dwarf honey bee
220	Pca1	954545135	540	9	*Polistes canadensis*	Red paper wasp
Cluster 12E			Average: 494 +/- 18			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
221	Obo1	952541846	500	9	*Oryctes borbonicus*	Rhinoceros beetle
222	Tca2	642923221	480	9	*Tribolium castaneum*	Red flour beetle
223	Dpo2	546673259	517	11	*Dendroctonus ponderosae*	Mountain pine beetle
224	Zne1	646693820	478	9	*Zootermopsis nevadensis*	Dampwood termite

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

Cluster 13			Average: 928 +/- 173			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
225	Olu1	145345985	884	7	*Ostreococcus lucimarinus CCE9901*	Green algae
226	Bpr1	612394653	1119	10	*Bathycoccus prasinos*	Picoplankton
227	Csu1	545373185	781	9	*Coccomyxa subellipsoidea C-169*	Microalgae (Variant)
Cluster 14			Average: 872 +/- 187			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
228	Tth1	829184078	1004	7	*Tetrahymena thermophila SB210*	No common name
229	Imu1	471225344	739	9	*Ichthyophthirius multifiliis*	Freshwater white spot disease
Cluster 15			Average: 825 +/- 203			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
230	Pxu4	910310360	844	9	*Papilio xuthus*	Asian swallowtail
231	Ppo2	909555601	591	11	*Papilio polytes*	Common Mormon

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

232	Bmo3	827545153	888	9	*Bombyx mori*	Silkworm
233	Api1	641651628	654	10	*Acyrthosiphon pisum*	Pea aphid
234	Zne2	646718108	806	11	*Zootermopsis nevadensis*	Dampwood termite
235	Ssc1	934157056	1167	8	*Sarcoptes scabiei*	Itch mite
Cluster 16			Average: 457 +/- 15			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name
236	Cel1	71986837	458	8	*Caenorhabditis elegans*	Roundworm
237	Cbr1	268581051	460	8	*Caenorhabditis briggsae*	No common name
238	Ace1	597854144	483	8	*Ancylostoma ceylanicum*	Roundworm (Variant)
239	Asu1	541045997	457	8	*Ascaris suum*	Large roundworm of pigs
240	Tca5	73454630	450	7	*Toxocara canis*	Dog roundworm
241	Bma1	671412333	436	8	*Brugia malayi*	Round worm (Variant)
Cluster 17			Average: 1130 +/- 93			
Protein Number (Clockwise)	Abbreviation	GI number	Number of amino acids	Number of predicted TMSs	*Organism*	Common name

Table 3: 248 KXA family homologues listed in clockwise order based on phylogenetic tree, Continued

242	Pxy3	768414960	1100	9	*Plutella xylostella*	Diamondback moth
243	Pxu2	910319838	1115	10	*Papilio xuthus*	Asian swallowtail
244	Bmo4	827562732	1105	10	*Bombyx mori*	Silkworm
245	Aro1	817062344	1258	10	*Athalia rosae*	Turnip sawfly
246	Oab2	817198111	1256	10	*Orussus abietinus*	Sawfly
247	Cfl2	752870900	1032	9	*Camponotus floridanus*	Florida carpenter ant
248	Zne4	646720184	1041		*Zootermopsis nevadensis*	Dampwood termite

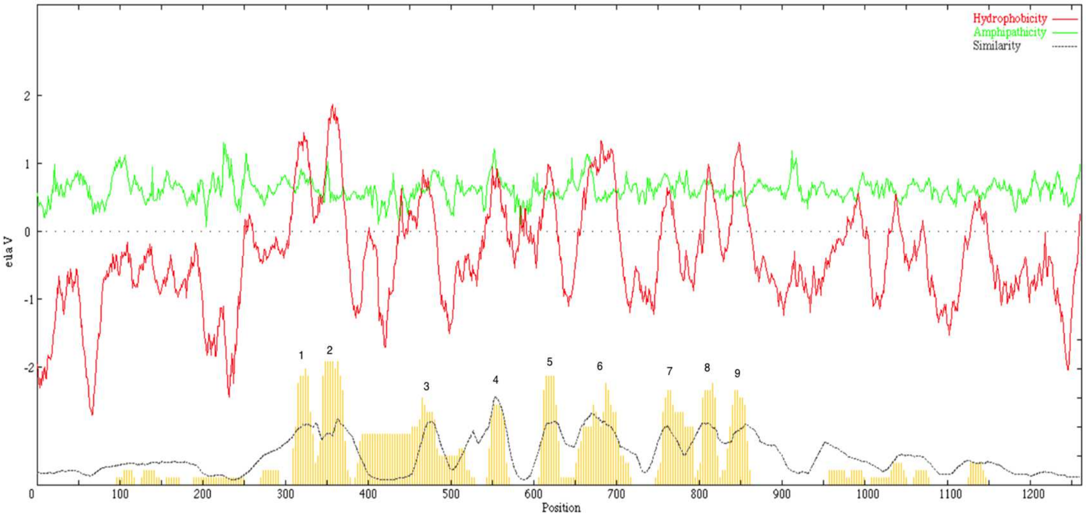


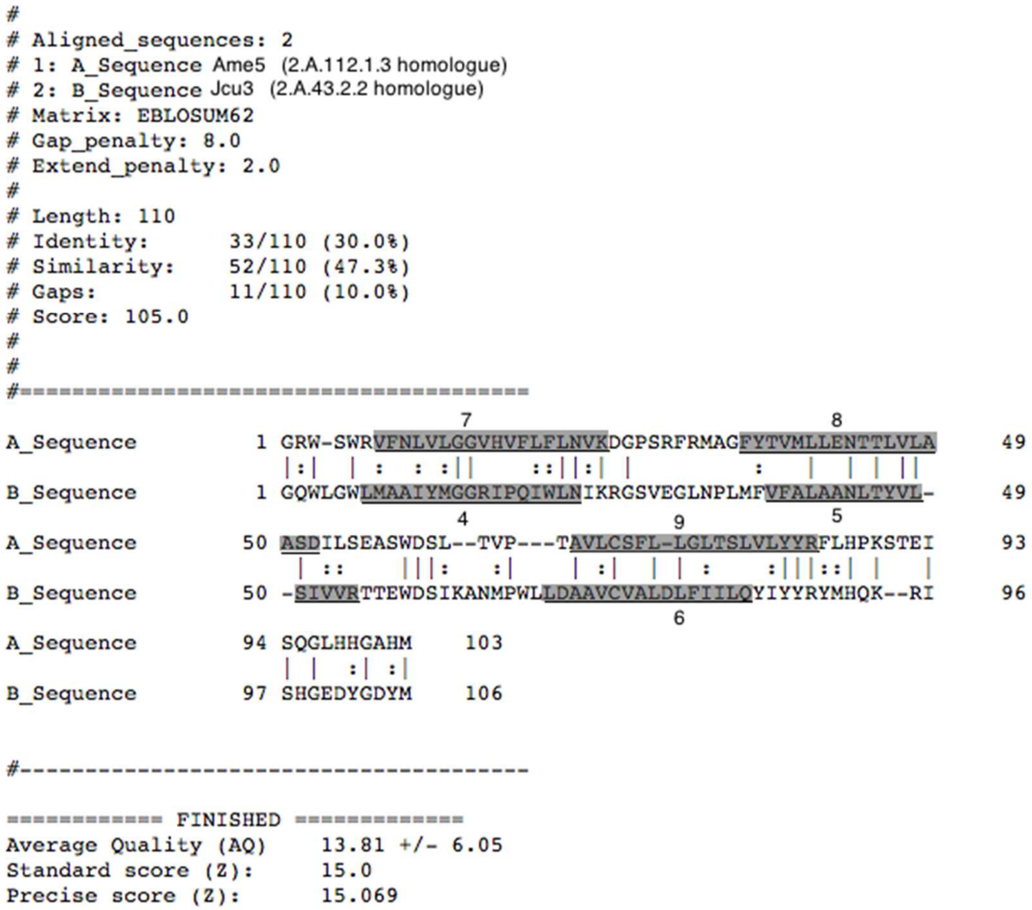
Figure 8: Average Hydropathy, Amphipathicity, and Similarity Plots (AveHAS) for all members of the KX Blood-group Antigen (KXA) Family included in TCDB as of date.

Table 4: Comparisons between the KX Blood-group Antigen (KXA) family and the Transporter-Opsin-G protein-coupled receptor (TOG) superfamily.

Families compared	Protein A	Protein B	Protein C	Protein D	A v B	B v C	C v D	A v D
XK v LCT	Q8C0T0	W5K2Y7	XP_012079100.1	Q95XZ6	31	15	34	0.1
XK v NiCoT	Q8C0T0	XP_012677872.1	A0A0E9NFW4	Q478Q8	45	15	21	0.34
XK v PNaS	P51811	XP_008273837.1	XP_015272244.1	Q06495	157	14	239	0.76
XK v OST	Q8C0T0	E2A6Z0	KFH63302.1	Q9Y519	51	17	60	0.9
XK v TSUP	Q5GH70	XP_014235795.1	A0A0F2RHA9	Q89V93	30	15	66	1.4
XK v Sweet	Q8C0T0	XP_004074365.1	XP_009783143.1	O96245	24	20	37	0.8
XK v MR	Q6UX68	A0A091H593	K8F6P9	Q5UXY6	150	16	18	0.3
XK v PnuC	E9FYJ5	A0A0B2UUP4	A0A090Y837	Q837U1	49	14	21	0.7

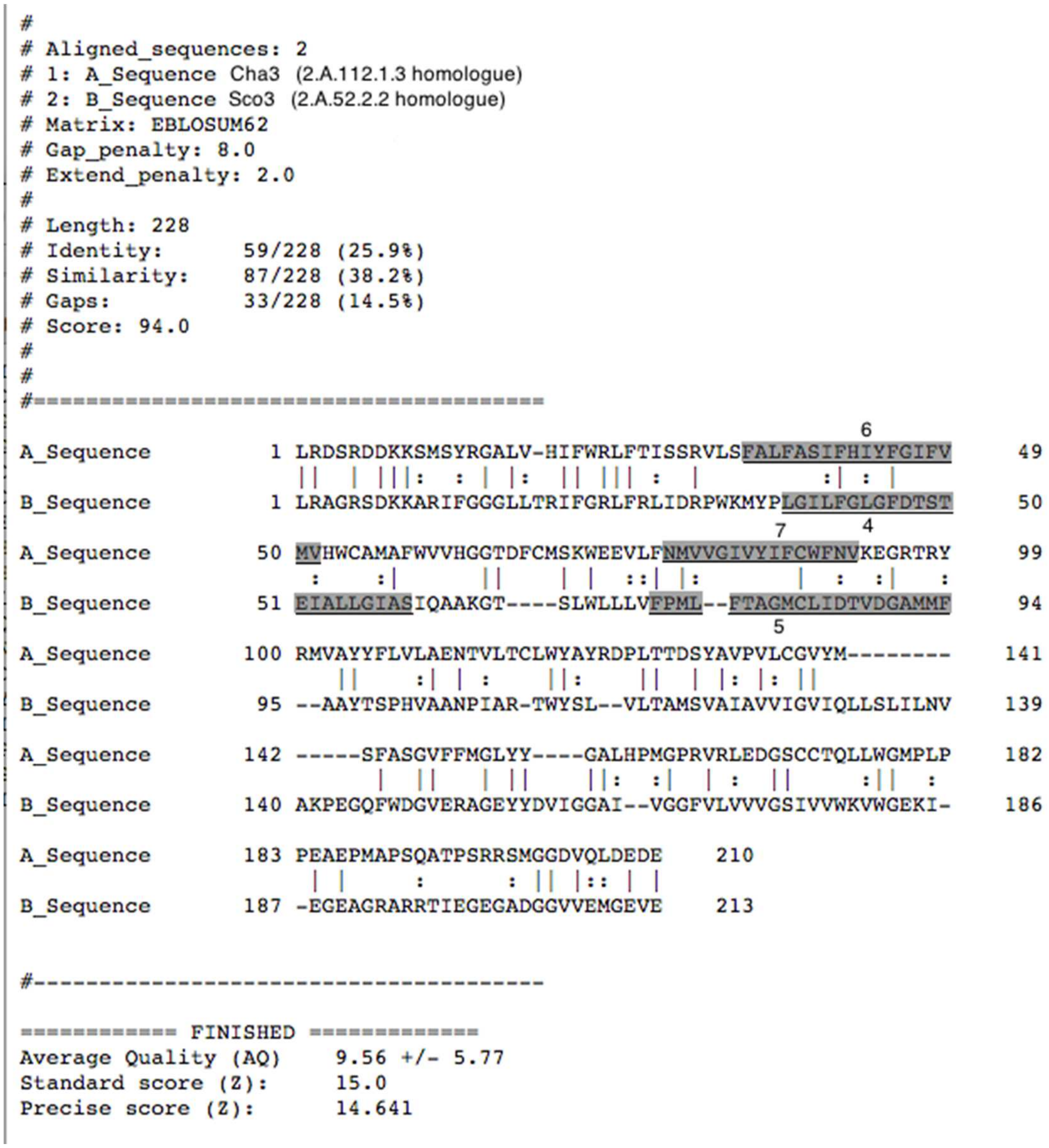
Table 4: The KX Blood-group Antigen (KXA) family and the Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily, continued.

XK v GPCR	Q8C0T0	XP_007061801.1	XP_014217479.1	Q9VKA4	37	15	88	0.7



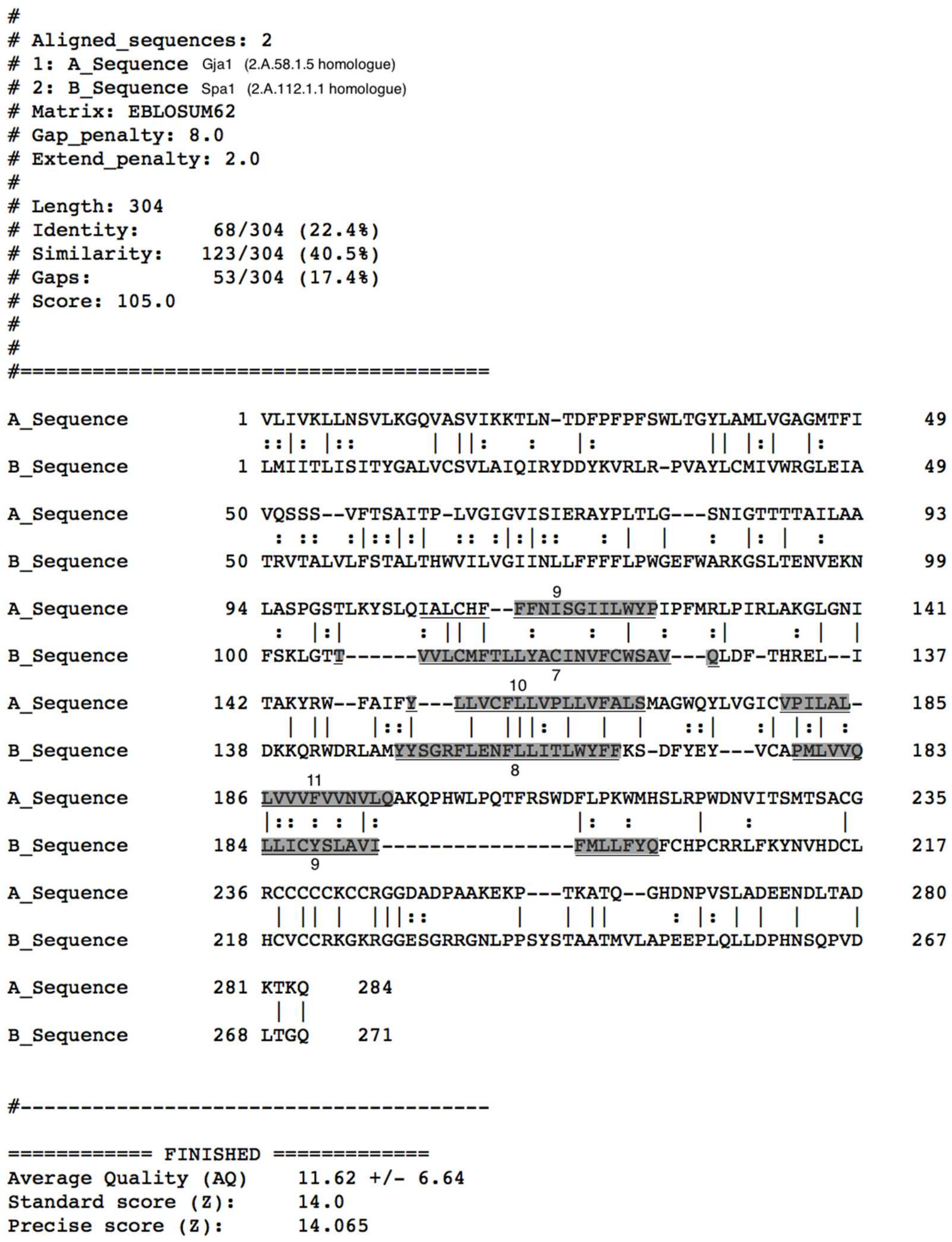
```
#
# Aligned_sequences: 2
# 1: A_Sequence Ame5  (2.A.112.1.3 homologue)
# 2: B_Sequence Jcu3  (2.A.43.2.2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 110
# Identity:      33/110 (30.0%)
# Similarity:    52/110 (47.3%)
# Gaps:          11/110 (10.0%)
# Score: 105.0
#
#
#=======================================

                                   7                        8
A_Sequence         1 GRW-SWRVFNLVLGGVHVFLFLNVKDGPSRFRMAGFYTVMLLENTTLVLA     49
                     |:|  | :  : :||    ::||:| |         :   |  | | ||
B_Sequence         1 GQWLGWLMAAIYMGGRIPQIWLNIKRGSVEGLNPLMFVFALAANLTYVL-     49
                                   4                9           5
A_Sequence        50 ASDILSEASWDSL--TVP---TAVLCSFL-LGLTSLVLYYRFLHPKSTEI     93
                      | ::    |||:   :|    | :|  | | :    :|||::| |   |
B_Sequence        50 -SIVVRTTEWDSIKANMPWLLDAAVCVALDLFIILQYIYYRYMHQK--RI     96
                                                    6
A_Sequence        94 SQGLHHGAHM    103
                     | |  :| :|
B_Sequence        97 SHGEDYGDYM    106


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     13.81 +/- 6.05
Standard score (Z):      15.0
Precise score (Z):       15.069
```

Figure 9: GSAT alignment comparing the KXA family homologue, Ame5, and the LCT family homologue, Jcu3.



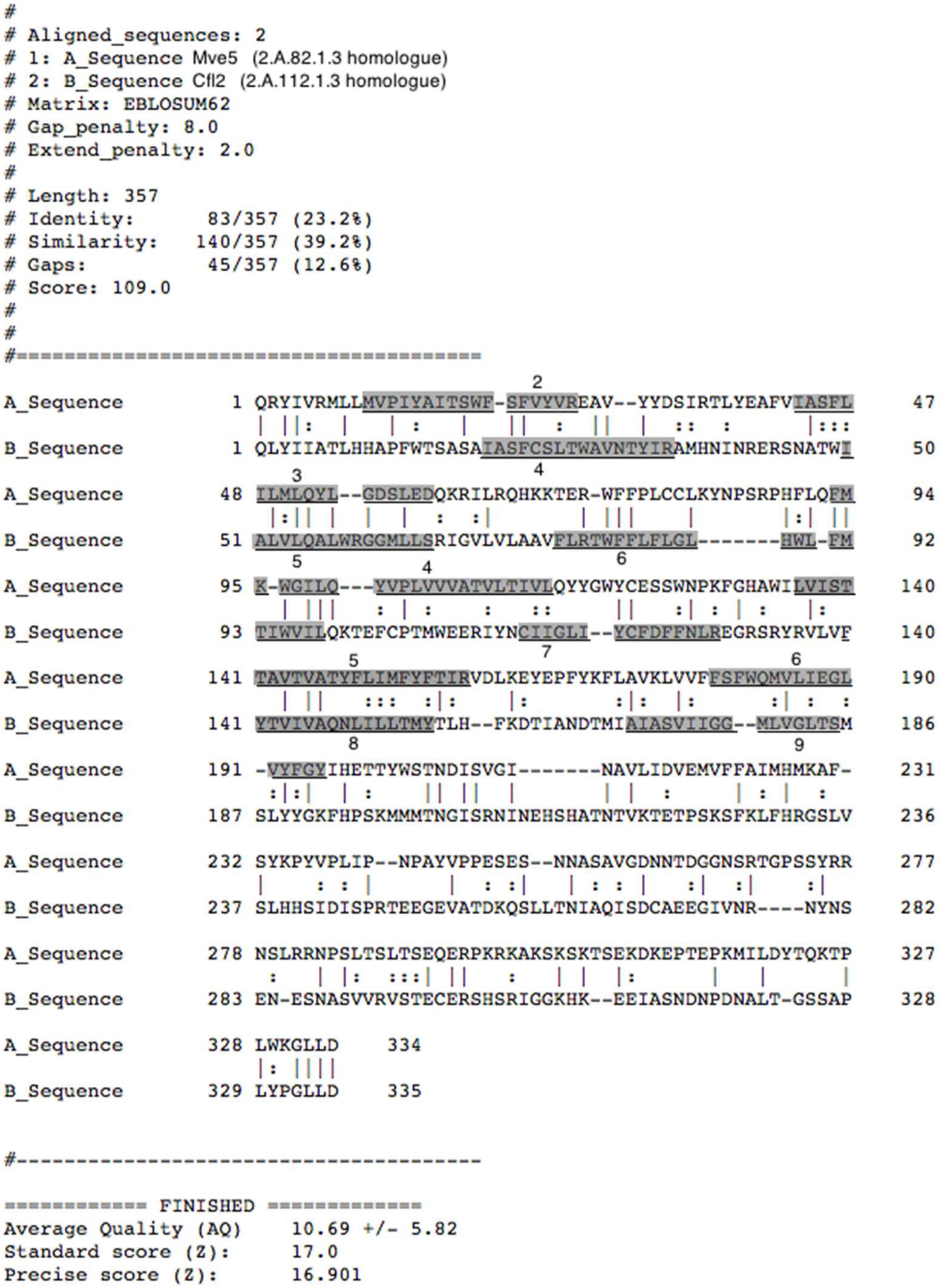
```
#
# Aligned_sequences: 2
# 1: A_Sequence Cha3  (2.A.112.1.3 homologue)
# 2: B_Sequence Sco3  (2.A.52.2.2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 228
# Identity:      59/228 (25.9%)
# Similarity:    87/228 (38.2%)
# Gaps:          33/228 (14.5%)
# Score: 94.0
#
#
#=======================================

                                                                6
A_Sequence         1 LRDSRDDKKSMSYRGALV-HIFWRLFTISSRVLSFALFASIFHIYFGIFV     49
                     ||  | |||:  : | |:  || ||| :  |         :| : |
B_Sequence         1 LRAGRSDKKARIFGGGLLTRIFGRLFRLIDRPWKMYPLGILFGLGFDTST     50
                                                          7     4
A_Sequence        50 MVHWCAMAFWVVHGGTDFCMSKWEEVLFNMVVGIVYIFCWFNVKEGRTRY     99
                      :    :|      ||    | |  ::| |:       |  :  :|   :
B_Sequence        51 EIALLGIASIQAAKGT----SLWLLLVFPML--FTAGMCLIDTVDGAMMF     94
                                                          5
A_Sequence       100 RMVAYYFLVLAENTVLTCLWYAYRDPLTTDSYAVPVLCGVYM--------    141
                        ||    :| | :    ||:    ||  | |: |: ||
B_Sequence        95 --AAYTSPHVAANPIAR-TWYSL--VLTAMSVAIAVVIGVIQLLSLILNV    139

A_Sequence       142 -----SFASGVFFMGLYY----GALHPMGPRVRLEDGSCCTQLLWGMPLP    182
                           |  ||   |  ||    ||:  :|  | :  ||     :||  :
B_Sequence       140 AKPEGQFWDGVERAGEYYDVIGGAI--VGGFVLVVVGSIVVWKVWGEKI-    186

A_Sequence       183 PEAEPMAPSQATPSRRSMGGDVQLDEDE    210
                      | |     :      : || |:: | |
B_Sequence       187 -EGEAGRARRTIEGEGADGGVVEMGEVE    213


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)    9.56 +/- 5.77
Standard score (Z):     15.0
Precise score (Z):      14.641
```

Figure 10: GSAT alignment comparing the KXA family homologue, Cha3, and the NiCoT family homologue, Sco3.



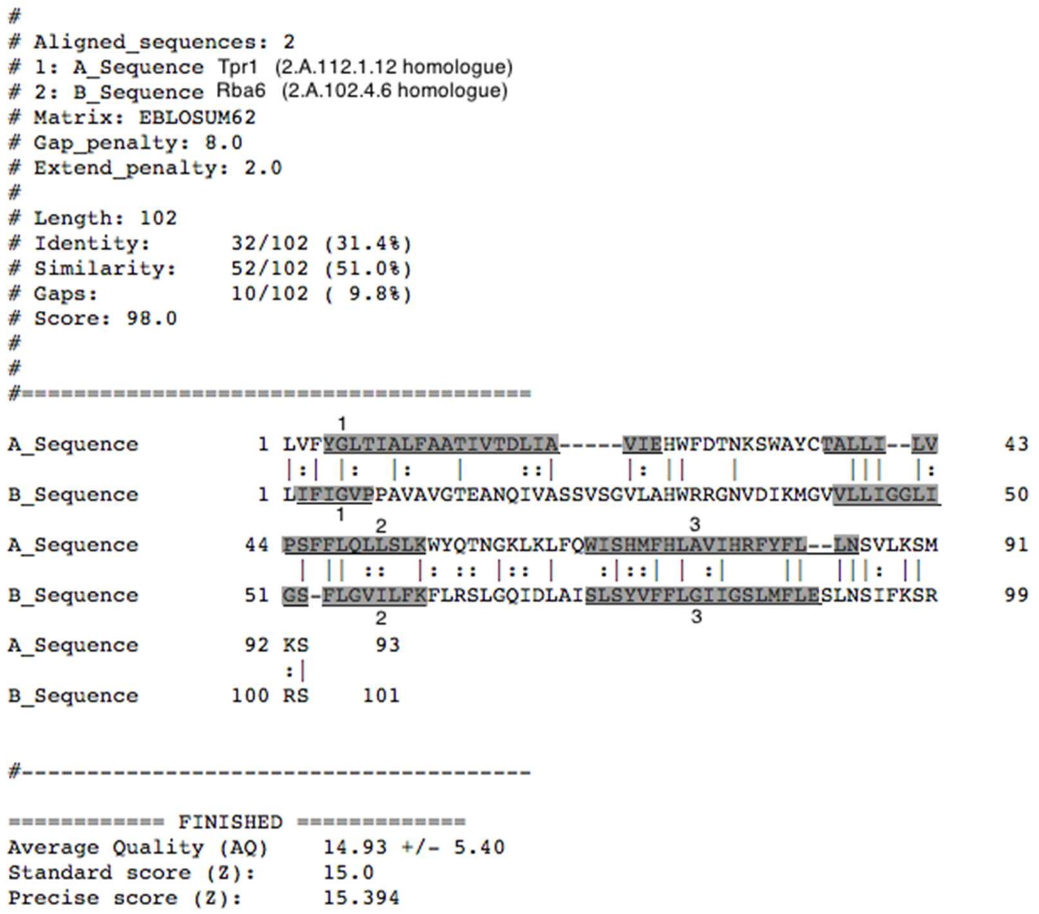
```
#
# Aligned_sequences: 2
# 1: A_Sequence Gja1  (2.A.58.1.5 homologue)
# 2: B_Sequence Spa1  (2.A.112.1.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 304
# Identity:      68/304 (22.4%)
# Similarity:   123/304 (40.5%)
# Gaps:          53/304 (17.4%)
# Score: 105.0
#
#
#=======================================

A_Sequence         1 VLIVKLLNSVLKGQVASVIKKTLN-TDFPFPFSWLTGYLAMLVGAGMTFI     49
                     ::|: |::      | ||:   :   |:         || |:|  |:
B_Sequence         1 LMIITLISITYGALVCSVLAIQIRYDDYKVRLR-PVAYLCMIVWRGLEIA     49

A_Sequence        50 VQSSS--VFTSAITP-LVGIGVISIERAYPLTLG---SNIGTTTTAILAA     93
                      : ::  :|::|:|  :: :|:|::   : |  |   :  |: |  :
B_Sequence        50 TRVTALVLFSTALTHWVILVGIINLLFFFFLPWGEFWARKGSLTENVEKN     99

A_Sequence        94 LASPGSTLKYSLQIALCHF--FFNISGIILWYPIPFMRLPIRLAKGLGNI    141
                      :  |:|      :  || |   :    :  |  :   :|     : |  |
B_Sequence       100 FSKLGTT------VVLCMFTLLYACINVFCWSAV---QLDF-THREL--I    137

A_Sequence       142 TAKYRW--FAIFY---LLVCFLLVPLLVFALSMAGWQYLVGICVPILAL-    185
                       | ||   |::|    |  |||: |  |  |   ::|   :| |:| :
B_Sequence       138 DKKQRWDRLAMYYSGRFLENFLLITLWYFFKS-DFYEY---VCAPMLVVQ    183

A_Sequence       186 LVVVFVVNVLQAKQPHWLPQTFRSWDFLPKWMHSLRPWDNVITSMTSACG    235
                     |:: : : |:                |:  :     |   :       |
B_Sequence       184 LLICYSLAVI----------------FMLLFYQFCHPCRRLFKYNVHDCL    217

A_Sequence       236 RCCCCCKCCRGGDADPAAKEKP---TKATQ--GHDNPVSLADEENDLTAD    280
                      | || |  |||::       |   | ||     : |: | |  |    |
B_Sequence       218 HCVCCRKGKRGGESGRRGNLPPSYSTAATMVLAPEEPLQLLDPHNSQPVD    267

A_Sequence       281 KTKQ    284
                      | |
B_Sequence       268 LTGQ    271


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)    11.62 +/- 6.64
Standard score (Z):     14.0
Precise score (Z):      14.065
```

Figure 11: GSAT alignment comparing the KXA family homologue, Spa1, and the PNaS family homologue, Gja1.



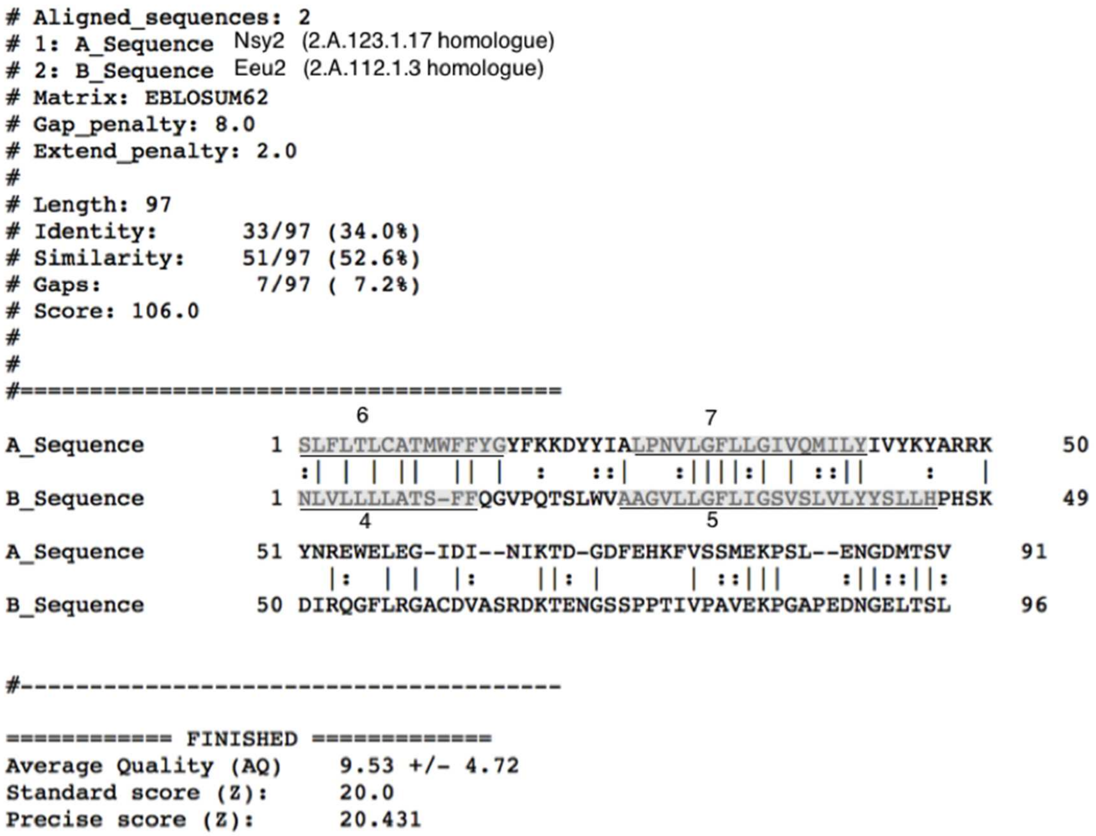
```
#
# Aligned_sequences: 2
# 1: A_Sequence Mve5  (2.A.82.1.3 homologue)
# 2: B_Sequence Cfl2  (2.A.112.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 357
# Identity:      83/357 (23.2%)
# Similarity:   140/357 (39.2%)
# Gaps:          45/357 (12.6%)
# Score: 109.0
#
#
#=======================================

                                                2
A_Sequence             1 QRYIVRMLLMVPIYAITSWF-SFVYVREAV--YYDSIRTLYEAFVIASFL     47
                         | ||:  |   | :   |   ||  :  ||  |  ::  :      |:::
B_Sequence             1 QLYIIATLHHAPFWTSASAIASFCSLTWAVNTYIRAMHNINRERSNATWI     50
                            3                   4
A_Sequence            48 ILMLQYL--GDSLEDQKRILRQHKKTER-WFFPLCCLKYNPSRPHFLQFM     94
                          |:|| |  |  |  :  :|       | |||    |       |:| ||
B_Sequence            51 ALVLQALWRGGMLLSRIGVLVLAAVFLRTWFFLFLGL-------HWL-FM     92
                            5          4               6
A_Sequence            95 K-WGILQ---YVPLVVVATVLTIVLQYYGWYCESSWNPKFGHAWILVIST    140
                           | |||   : | :    :   ::     ||   :| : | :   |:
B_Sequence            93 TIWVILQKTEFCPTMWEERIYNCIIGLI--YCFDFFNLREGRSRYRVLVF    140
                                                 7
                                 5                                    6
A_Sequence           141 TAVTVATYFLIMFYFTIRVDLKEYEPFYKFLAVKLVVFFSFWQMVLIEGL    190
                           | ||   :::  :|:    |:       :|:  |:      :| :  :
B_Sequence           141 YTVIVAQNLILLTMYTLH--FKDTIANDTMIAIASVIIGG--MLVGLTSM    186
                                 8                                    9
A_Sequence           191 -VYFGYIHETTYWSTNDISVGI-------NAVLIDVEMVFFAIMHMKAF-    231
                          :|:|  | :    || ||  |       | |  :     | : |  :
B_Sequence           187 SLYYGKFHPSKMMMTNGISRNINEHSHATNTVKTETPSKSFKLFHRGSLV    236

A_Sequence           232 SYKPYVPLIP--NPAYVPPESES--NNASAVGDNNTDGGNSRTGPSSYRR    277
                         |    : : |      |  : :|   | : : |   :|  :|    :|
B_Sequence           237 SLHHSIDISPRTEEGEVATDKQSLLTNIAQISDCAEEGIVNR----NYNS    282

A_Sequence           278 NSLRRNPSLTSLTSEQERPKRKAKSKSKTSEKDKEPTEPKMILDYTQKTP    327
                          :   | |:  :::| ||   :   | |  |:      |   |      |
B_Sequence           283 EN-ESNASVVRVSTECERSHSRIGGKHK--EEIASNDNPDNALT-GSSAP    328

A_Sequence           328 LWKGLLD     334
                         |: ||||
B_Sequence           329 LYPGLLD     335


#---------------------------------------

============ FINISHED ============
Average Quality (AQ)    10.69 +/- 5.82
Standard score (Z):     17.0
Precise score (Z):      16.901
```

Figure 12: GSAT alignment comparing the KXA family homologue, Cfl2, and the OST family homologue, Mve5.



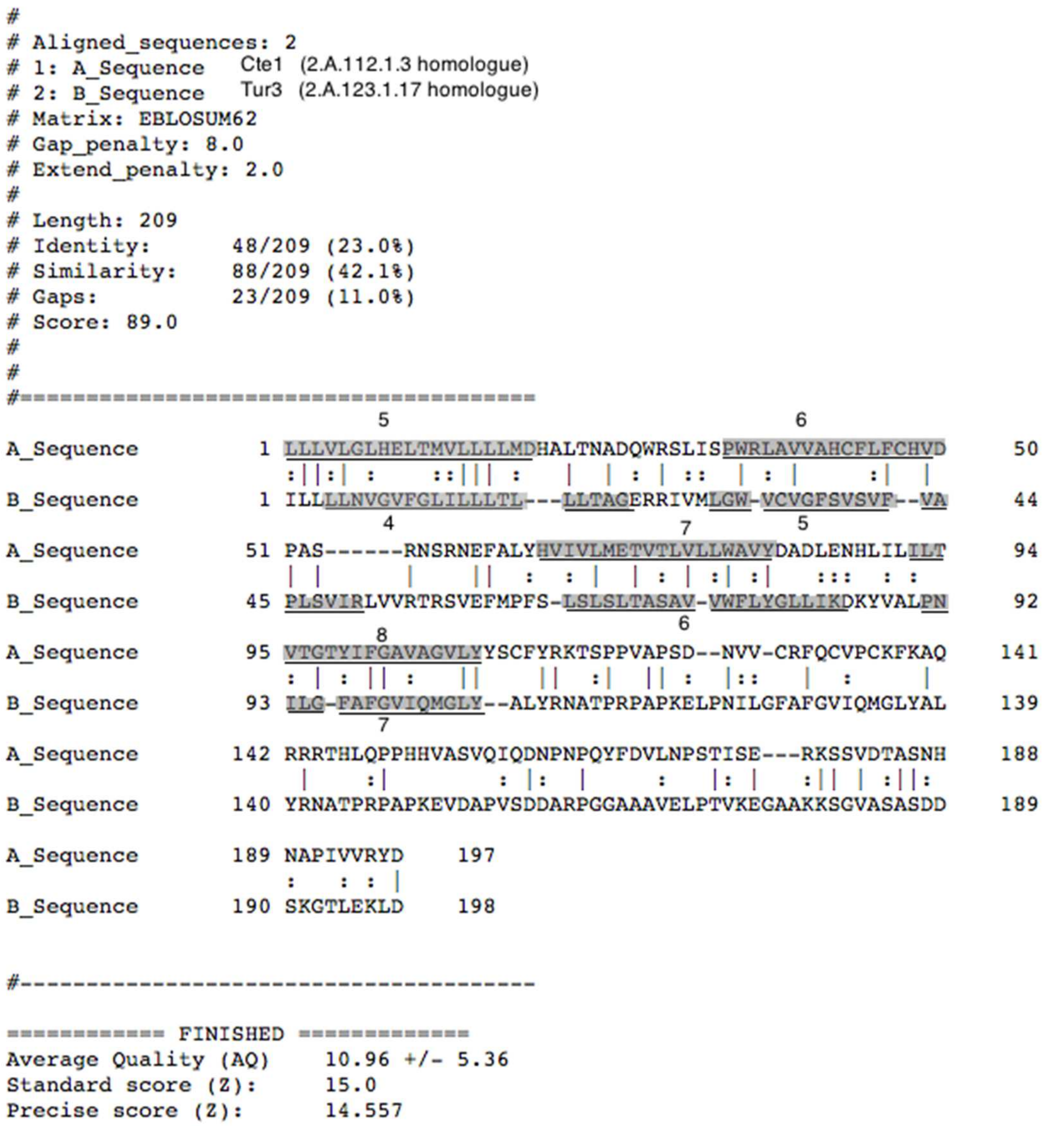
```
#
# Aligned_sequences: 2
# 1: A_Sequence Tpr1  (2.A.112.1.12 homologue)
# 2: B_Sequence Rba6  (2.A.102.4.6 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 102
# Identity:      32/102 (31.4%)
# Similarity:    52/102 (51.0%)
# Gaps:          10/102 ( 9.8%)
# Score: 98.0
#
#
#=======================================

                            1
A_Sequence         1 LVFYGLTIALFAATIVTDLIA-----VIEHWFDTNKSWAYCTALLI--LV     43
                     |:| |:  |:   |    ::|     |: ||   |        |||  |:
B_Sequence         1 LIFIGVPPAVAVGTEANQIVASSVSGVLAHWRRGNVDIKMGVVLLIGGLI     50
                         1  2                       3
A_Sequence        44 PSFFLQLLSLKWYQTNGKLKLFQWISHMFHLAVIHRFYFL--LNSVLKSM     91
                      | || ::  |: :: |:: |   :|::| | :|    ||  |||: ||
B_Sequence        51 GS-FLGVILFKFLRSLGQIDLAISLSYVFFLGIIGSLMFLESLNSIFKSR     99
                            2                       3
A_Sequence        92 KS     93
                     :|
B_Sequence       100 RS    101


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     14.93 +/- 5.40
Standard score (Z):      15.0
Precise score (Z):       15.394
```

Figure 13: GSAT alignment comparing the KXA family homologue, Tpr1, and the TSUP family homologue, Rba6.



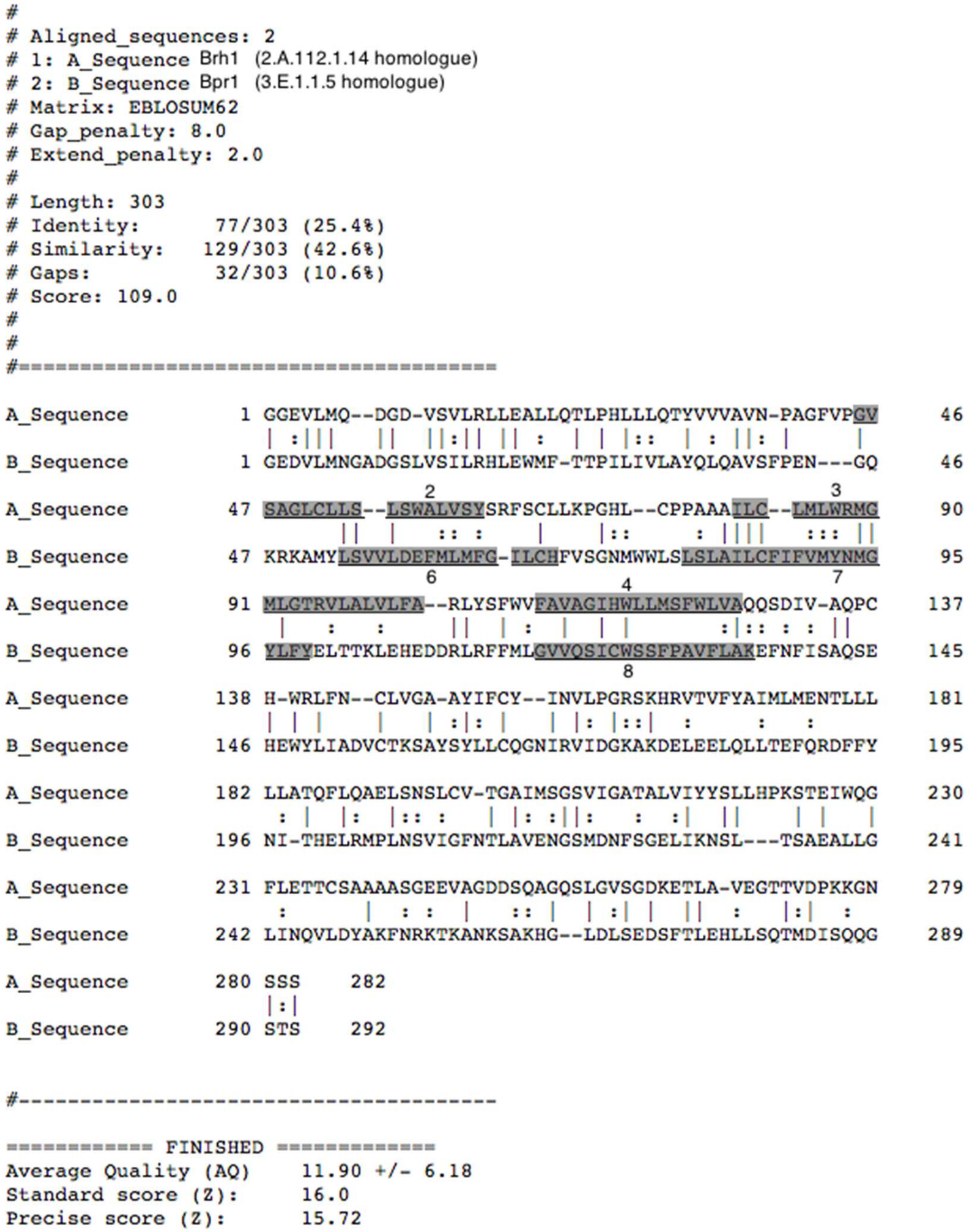
```
# Aligned_sequences: 2
# 1: A_Sequence Nsy2  (2.A.123.1.17 homologue)
# 2: B_Sequence Eeu2  (2.A.112.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 97
# Identity:      33/97 (34.0%)
# Similarity:    51/97 (52.6%)
# Gaps:           7/97 ( 7.2%)
# Score: 106.0
#
#
#=======================================
                         6                        7
A_Sequence         1 SLFLTLCATMWFFYGYFKKDYYIALPNVLGFLLGIVQMILYIVYKYARRK     50
                     :| | | ||  || |  :   ::|   :||||:| | ::||    :   |
B_Sequence         1 NLVLLLLATS-FFQGVPQTSLWVAAGVLLGFLIGSVSLVLYYSLLHPHSK     49
                         4                        5
A_Sequence        51 YNREWELEG-IDI--NIKTD-GDFEHKFVSSMEKPSL--ENGDMTSV     91
                       |:  | |  |:    ||: |      | ::|||    :||::||:
B_Sequence        50 DIRQGFLRGACDVASRDKTENGSSPPTIVPAVEKPGAPEDNGELTSL     96


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     9.53 +/- 4.72
Standard score (Z):      20.0
Precise score (Z):       20.431
```

Figure 14a: GSAT alignment comparing the KXA family homologue, Eeu2, and the Sweet family homologue, Nsy2.



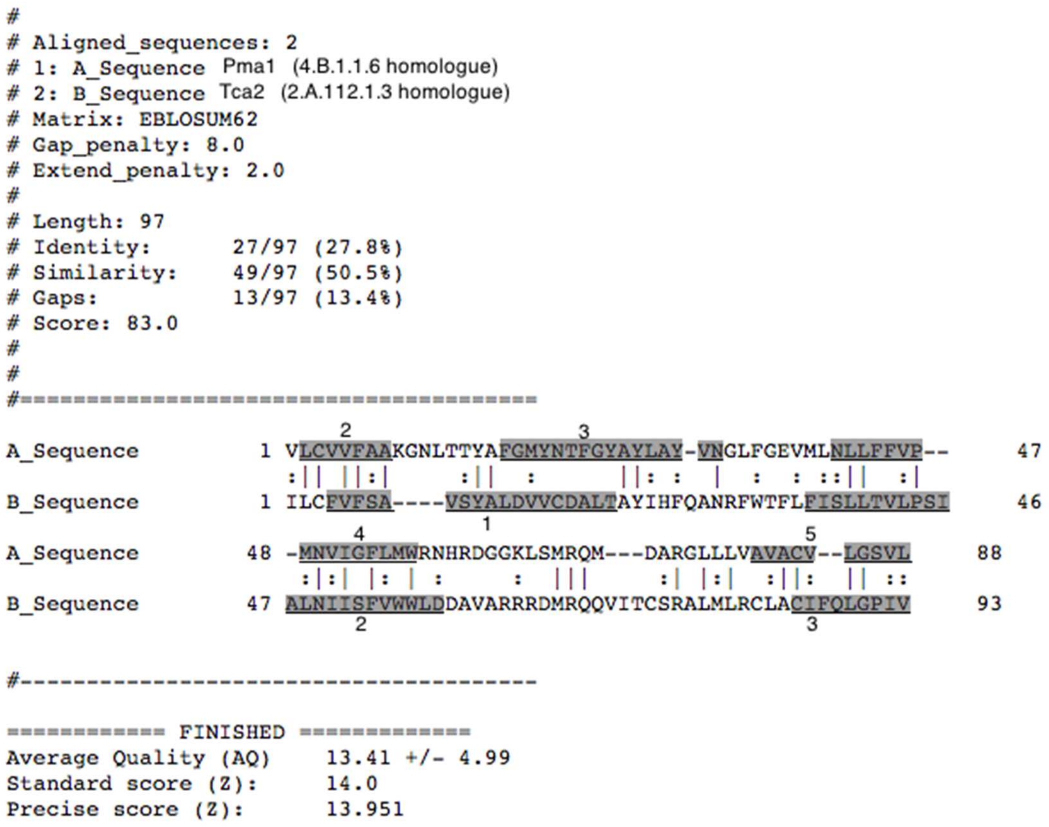
```
#
# Aligned_sequences: 2
# 1: A_Sequence   Cte1  (2.A.112.1.3 homologue)
# 2: B_Sequence   Tur3  (2.A.123.1.17 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 209
# Identity:      48/209 (23.0%)
# Similarity:    88/209 (42.1%)
# Gaps:          23/209 (11.0%)
# Score: 89.0
#
#
#=======================================
                            5                              6
A_Sequence          1 LLLVLGLHELTMVLLLLMDHALTNADQWRSLISPWRLAVVAHCFLFCHVD     50
                      :||:| :    ::||| :   |  | : | ::  | : |     :|  |
B_Sequence          1 ILLLLNVGVFGLILLLTL---LLTAGERRIVMLGW-VCVGFSVSVF--VA     44
                              4                     7       5
A_Sequence         51 PAS------RNSRNEFALYHVIVLMETVTLVLLWAVYDADLENHLILILT     94
                      | |      |    ||  :  : |  | : | :| :|   :::  : :
B_Sequence         45 PLSVIRLVVRTRSVEFMPFS-LSLSLTASAV-VWFLYGLLIKDKYVALPN     92
                                                    6
                             8
A_Sequence         95 VTGTYIFGAVAGVLYYSCFYRKTSPPVAPSD--NVV-CRFQCVPCKFKAQ    141
                      : | : || :   ||    ||  :|  || :  |::   |  :     |
B_Sequence         93 ILG-FAFGVIQMGLY--ALYRNATPRPAPKELPNILGFAFGVIQMGLYAL    139
                             7
A_Sequence        142 RRRTHLQPPHHVASVQIQDNPNPQYFDVLNPSTISE---RKSSVDTASNH    188
                       |    :|        : |:  |     :   |: |   :|| | :||:
B_Sequence        140 YRNATPRPAPKEVDAPVSDDARPGGAAAVELPTVKEGAAKKSGVASASDD    189

A_Sequence        189 NAPIVVRYD    197
                      :   : :|
B_Sequence        190 SKGTLEKLD    198


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     10.96 +/- 5.36
Standard score (Z):      15.0
Precise score (Z):       14.557
```

Figure 14b: GSAT alignment comparing the KXA family homologue, Cte1, and the Sweet family homologue, Tur3.



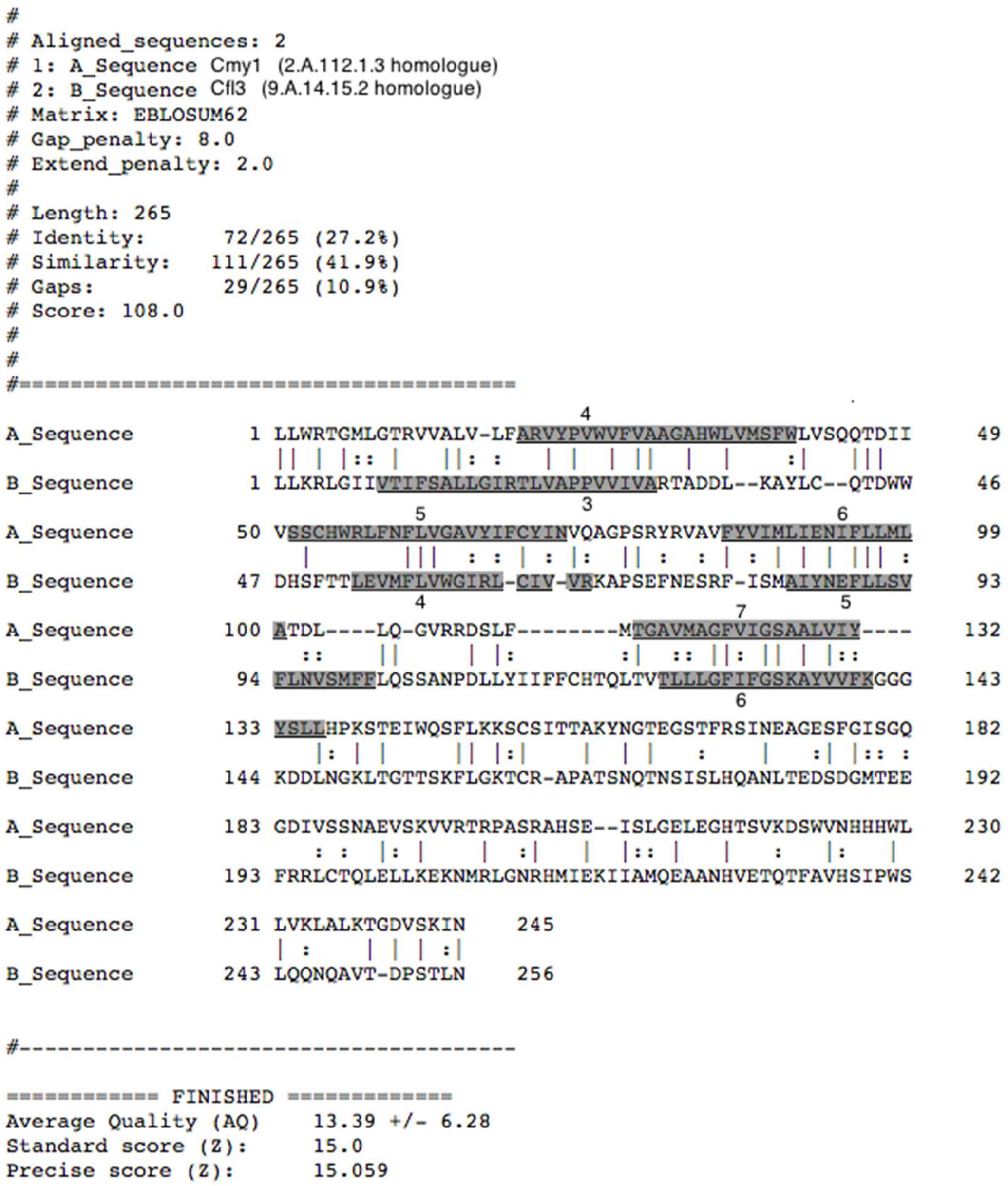
```
#
# Aligned_sequences: 2
# 1: A_Sequence Brh1  (2.A.112.1.14 homologue)
# 2: B_Sequence Bpr1  (3.E.1.1.5 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 303
# Identity:      77/303 (25.4%)
# Similarity:   129/303 (42.6%)
# Gaps:          32/303 (10.6%)
# Score: 109.0
#
#
#=======================================

A_Sequence         1 GGEVLMQ--DGD-VSVLRLLEALLQTLPHLLLQTYVVVAVN-PAGFVPGV     46
                     | :|||   ||  ||:|| || :  | | |::  | : ||: |     |
B_Sequence         1 GEDVLMNGADGSLVSILRHLEWMF-TTPILIVLAYQLQAVSFPEN---GQ     46
                                  2                                3
A_Sequence        47 SAGLCLLS--LSWALVSYSRFSCLLKPGHL--CPPAAAILC--LMLWRMG     90
                           ||  |   :: :    |    |::     : ||||   ::: ||
B_Sequence        47 KRKAMYLSVVLDEFMLMFG-ILCHFVSGNMWWLSLSLAILCFIFVMYNMG     95
                                  6                                7
                                                  4
A_Sequence        91 MLGTRVLALVLFA--RLYSFWVFAVAGIHWLLMSFWLVAQQSDIV-AQPC    137
                      |   :   :     ||  | :  |  | |       :|:: : : ||
B_Sequence        96 YLFYELTTKLEHEDDRLRFFMLGVVQSICWSSFPAVFLAKEFNFISAQSE    145
                                                  8
A_Sequence       138 H-WRLFN--CLVGA-AYIFCY--INVLPGRSKHRVTVFYAIMLMENTLLL    181
                     | | |    |   | :|: |   | |: |::|  :     :   :
B_Sequence       146 HEWYLIADVCTKSAYSYLLCQGNIRVIDGKAKDELEELQLLTEFQRDFFY    195

A_Sequence       182 LLATQFLQAELSNSLCV-TGAIMSGSVIGATALVIYYSLLHPKSTEIWQG    230
                      : |  |:  |:: :   | |: :||:   :  :|  ||    | |   |
B_Sequence       196 NI-THELRMPLNSVIGFNTLAVENGSMDNFSGELIKNSL---TSAEALLG    241

A_Sequence       231 FLETTCSAAAASGEEVAGDDSQAGQSLGVSGDKETLA-VEGTTVDPKKGN    279
                      :      |  : :  |   :: |  | :| |  ||  :   |:|  :
B_Sequence       242 LINQVLDYAKFNRKTKANKSAKHG--LDLSEDSFTLEHLLSQTMDISQQG    289

A_Sequence       280 SSS    282
                     |:|
B_Sequence       290 STS    292


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)    11.90 +/- 6.18
Standard score (Z):     16.0
Precise score (Z):      15.72
```

Figure 15: GSAT alignment comparing the KXA family homologue, Bpr1, and the MR family homologue, Brh1.



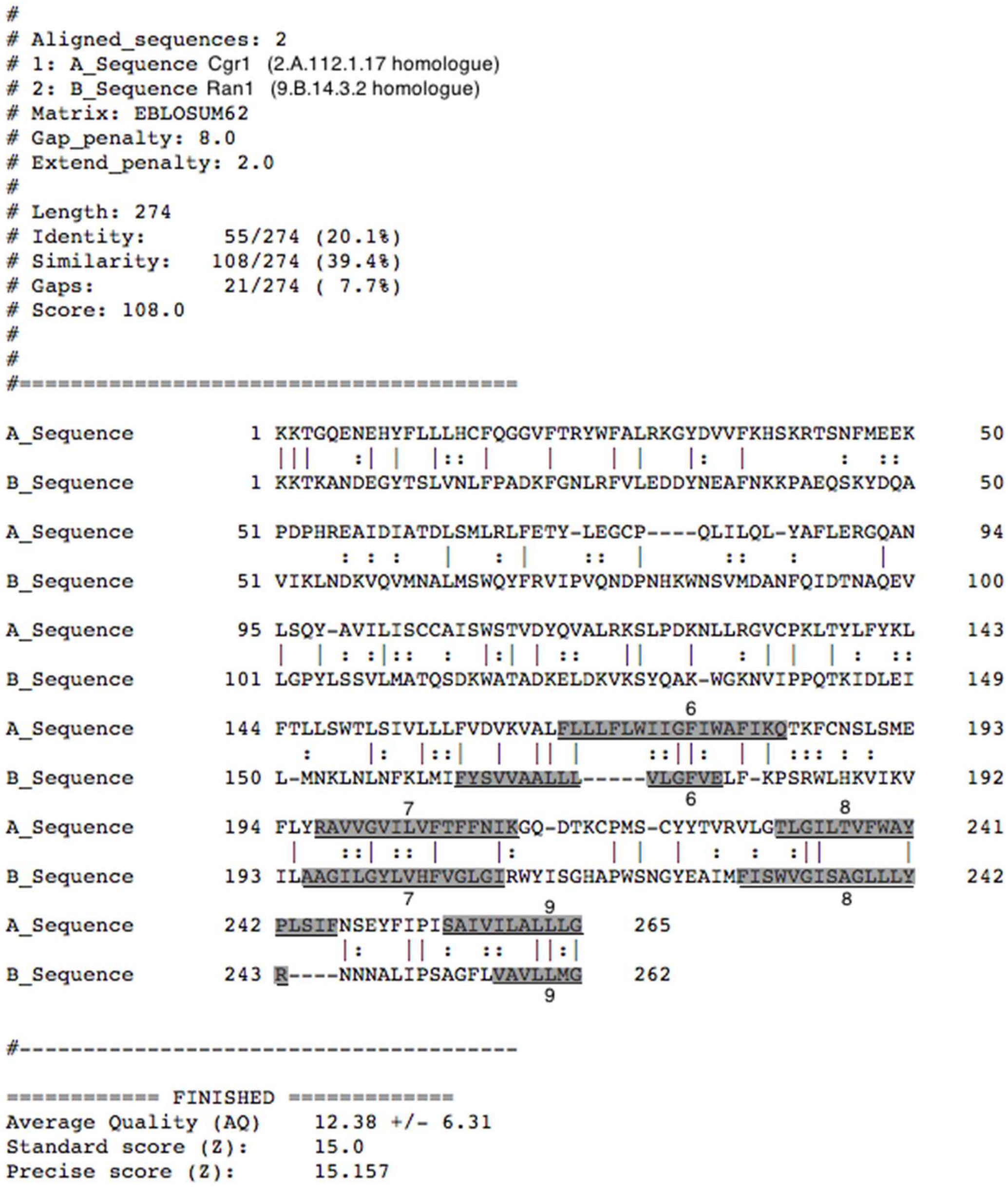
```
#
# Aligned_sequences: 2
# 1: A_Sequence Pma1  (4.B.1.1.6 homologue)
# 2: B_Sequence Tca2  (2.A.112.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 97
# Identity:      27/97 (27.8%)
# Similarity:    49/97 (50.5%)
# Gaps:          13/97 (13.4%)
# Score: 83.0
#
#
#=======================================

                          2                 3
A_Sequence          1 VLCVVFAAKGNLTTYAFGMYNTFGYAYLAY-VNGLFGEVMLNLLFFVP--     47
                      :|| ||:|     :||  :      ||: :  |  :  : ::||  :|
B_Sequence          1 ILCFVFSA----VSYALDVVCDALTAYIHFQANRFWTFLFISLLTVLPSI     46
                                     1
                           4                                 5
A_Sequence         48 -MNVIGFLMWRNHRDGGKLSMRQM---DARGLLLVAVACV--LGSVL     88
                       :|:| |: | :     :  |||     :| |:|  :||:  || ::
B_Sequence         47 ALNIISFVWWLDDAVARRRDMRQQVITCSRALMLRCLACIFQLGPIV     93
                           2                                 3

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     13.41 +/- 4.99
Standard score (Z):      14.0
Precise score (Z):       13.951
```

Figure 16: GSAT alignment comparing the KXA family homologue, Tca2, and the PnuC family homologue, Pma1.



```
#
# Aligned_sequences: 2
# 1: A_Sequence Cmy1  (2.A.112.1.3 homologue)
# 2: B_Sequence Cfl3  (9.A.14.15.2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 265
# Identity:      72/265 (27.2%)
# Similarity:   111/265 (41.9%)
# Gaps:          29/265 (10.9%)
# Score: 108.0
#
#
#=======================================

                                             4
A_Sequence         1 LLWRTGMLGTRVVALV-LFARVYPVWVFVAAGAHWLVMSFWLVSQQTDII      49
                     || | |:: |   ||: :   | |  | ||  |  |    :|   |||
B_Sequence         1 LLKRLGIIVTIFSALLGIRTLVAPPVVIVARTADDL--KAYLC--QTDWW      46
                                             3
                                5                                6
A_Sequence        50 VSSCHWRLFNFLVGAVYIFCYINVQAGPSRYRVAVFYVIMLIENIFLLML      99
                       |       ||| : : | : |:  || :  : | : | | | ||| :
B_Sequence        47 DHSFTTLEVMFLVWGIRL-CIV-VRKAPSEFNESRF-ISMAIYNEFLLSV      93
                                4                                5
                                                         7
A_Sequence       100 ATDL----LQ-GVRRDSLF--------MTGAVMAGFVIGSAALVIY----     132
                       ::    ||     | |:        :|  :: ||: || | |::
B_Sequence        94 FLNVSMFFLQSSANPDLLYIIFFCHTQLTVTLLLGFIFGSKAYVVFKGGG     143
                                                         6
A_Sequence       133 YSLLHPKSTEIWQSFLKKSCSITTAKYNGTEGSTFRSINEAGESFGISGQ     182
                        |: | |     || |:|    |  | |   :    |   :| |:: :
B_Sequence       144 KDDLNGKLTGTTSKFLGKTCR-APATSNQTNSISLHQANLTEDSDGMTEE     192

A_Sequence       183 GDIVSSNAEVSKVVRTRPASRAHSE--ISLGELEGHTSVKDSWVNHHHWL     230
                        : :  |: |    |  :|   |  |:: |   |   :   |:   |
B_Sequence       193 FRRLCTQLELLKEKNMRLGNRHMIEKIIAMQEAANHVETQTFAVHSIPWS     242

A_Sequence       231 LVKLALKTGDVSKIN    245
                     | :    | | |:|
B_Sequence       243 LQQNQAVT-DPSTLN    256


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)    13.39 +/- 6.28
Standard score (Z):     15.0
Precise score (Z):      15.059
```

Figure 17: GSAT alignment comparing the KXA family homologue, Cmy1, and the GPCR family homologue, Cfl3.



```
#
# Aligned_sequences: 2
# 1: A_Sequence Cgr1  (2.A.112.1.17 homologue)
# 2: B_Sequence Ran1  (9.B.14.3.2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 274
# Identity:      55/274 (20.1%)
# Similarity:   108/274 (39.4%)
# Gaps:          21/274 ( 7.7%)
# Score: 108.0
#
#
#=======================================

A_Sequence         1 KKTGQENEHYFLLLHCFQGGVFTRYWFALRKGYDVVFKHSKRTSNFMEEK     50
                     |||   :| |  |:: |    |    | |   |:  |       :  ::
B_Sequence         1 KKTKANDEGYTSLVNLFPADKFGNLRFVLEDDYNEAFNKKPAEQSKYDQA     50

A_Sequence        51 PDPHREAIDIATDLSMLRLFETY-LEGCP----QLILQL-YAFLERGQAN     94
                          : : :   |   : |    ::  |      ::   :      |
B_Sequence        51 VIKLNDKVQVMNALMSWQYFRVIPVQNDPNHKWNSVMDANFQIDTNAQEV    100

A_Sequence        95 LSQY-AVILISCCAISWSTVDYQVALRKSLPDKNLLRGVCPKLTYLFYKL    143
                     |  | : :|::  :  |:| | ::   ||   |   : | |  | :  ::
B_Sequence       101 LGPYLSSVLMATQSDKWATADKELDKVKSYQAK-WGKNVIPPQTKIDLEI    149
                                                     6
A_Sequence       144 FTLLSWTLSIVLLLFVDVKVALFLLLFLWIIGFIWAFIKQTKFCNSLSME    193
                       :    |:  |::|  |  || |     ::||:  | | ::: : :
B_Sequence       150 L-MNKLNLNFKLMIFYSVVAALLL-----VLGFVELF-KPSRWLHKVIKV    192
                               7                     6           8
A_Sequence       194 FLYRAVVGVILVFTFFNIKGQ-DTKCPMS-CYYTVRVLGTLGILTVFWAY    241
                      |   ::| :: |    |:       | |  |  :  :  :||      |
B_Sequence       193 ILAAGILGYLVHFVGLGIRWYISGHAPWSNGYEAIMFISWVGISAGLLLY    242
                               7          9                      8
A_Sequence       242 PLSIFNSEYFIPISAIVILALLLG     265
                          |:   || :  ::  ||:|
B_Sequence       243 R----NNNALIPSAGFLVAVLLMG     262
                                          9

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)    12.38 +/- 6.31
Standard score (Z):     15.0
Precise score (Z):      15.157
```

Figure 18: GSAT alignment comparing the KXA family homologue, Cgr1, and the HHP family homologue, Ran1.

References

1. Allen F.H. Jr., Krabbe S.M. and Corcoran P.A. (1961) A new phenotype (Mcleod) in the Kell blood-group system. Vox Sang. 6:555-60.

2. Baysse, C., S. Matthijs, M. Schobert, G. Layer, D. Jahn, and P. Cornelis. (2003). Co-ordination of iron acquisition, iron porphyrin chelation and iron-protoporphyrin export via the cytochrome c biogenesis protein CcmC in *Pseudomonas fluorescens.* Microbiology 149: 3543-3552.

3. Chen L.Q., Hou B.H., Lalonde S., Takanaga H., Hartung M.L., Qu X.Q., Guo W.J., Kim J.G., Underwood W., Chaudhuri B., Chermak D., Antony G., White F.F., Somerville S.C., Mudgett M.B. and Frommer W.B. (2010) Sugar transporters for intercellular exchange and nutrition of pathogens. Nature. 468(7323):527-32.

4. Danek A., Jung H.H., Melone M.A., Rampoldi L., Broccoli V. and Walker R.H. (2005) Neuroacanthocytosis: new developments in a neglected group of dementing disorders. J Neurol Sci. 229-230:171-86. Epub 2005 Jan 7.

5. Foster, J.W., Y.K. Park, T. Penfound, T. Fenger, and M.P. Spector. (1990). Regulation of NAD metabolism in *Salmonella typhimurium*: Molecular sequence analysis of the bifunctional *nadR* regulator and the *nadA-pnuC* operon. J. Bacteriol. 172: 4187-4196.

6. Frey B.M., Gassner C. and Jung H.H. (2015) Neurodegeneration in the elderly - When the blood type matters: An overview of the McLeod syndrome with focus on hematological features. Transfus Apher Sci. 52(3):277-84.

7. Hesse M., Pacini E. and Willemse M. (2004) The Tapetum: Cytology, Function, Biochemistry and Evolution. Springer, 2004. Pp. 1-152.

8. Ho M., Chelly J., Carter N., Danek A., Crocker P. and Monaco A.P. (1994) Isolation of the gene for McLeod syndrome that encodes a novel membrane transport protein. Cell. 77(6):869-80.

9. Jung H.H., Danek A. and Frey B.M. (2007) McLeod syndrome: a neurohaematological disorder. Vox Sang. 93(2): 112-121.

10. Kolbe, M., H. Besir, L. Essen, and D. Oesterhelt. (2000). Structure of the light-driven chloride pump halorhodopsin at 1.8 Å resolution. Science 288: 1390-1396.

11. Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. and Higgins D.G. (2007) Bioinformatics. 23(21):2947-8. Epub 2007 Sep 10.

12. Lee S., Zambas E.D. and Marsh W.L. and Redman C.M (1991) Molecular cloning and primary structure of the Kell blood group protein. Proc Natl Acad Sci U S A. 88(14):6353-7.

13. Lee S., Lin M., Mele A., Cao Y., Farmar J., Russo D. and Redman C. (1999) Proteolytic processing of big endothelin-3 by the Kell blood group protein. Blood. 94(4):1440-50.

14. Lee S., Sha Q., Wu X., Calenda G. and Peng J. (2007) Expression profiles of mouse Kell, XK, and XPLAC mRNA. J Histochem Cytochem. 55(4):365-74.

15. Lin I.W., Sosso D., Chen L.Q., Gase K., Kim S.G., Kessler D., Klinkenberg P.M., Gorder M.K., Hou B.H., Qu X.Q., Carter C.J., Baldwin I.T. and Frommer W.B. (2014) Nectar secretion requires sucrose phosphate synthases and the sugar transporter

SWEET9. Science. 508(7497):546-9.

16. Luecke, H., B. Schobert, H.-T. Richter, J.-P. Cartailler, and J.K. Lanyi. (1999). Structural changes in bacteriorhodopsin during ion transport at 2 angstrom resolution. Science 286: 255-260.

17. Mampel, J., E. Maier, T. Tralau, J. Ruff, R. Benz, and A.M. Cook. (2004). A novel outer-membrane anion channel (porin) as part of a putatively two-component transport system for 4-toluenesulphonate in *Comamonas testosteroni* T-2. Biochem. J. 383: 91-99.

18. Mayer, J., K. Denger, K. Hollemeyer, D. Schleheck, and A.M. Cook. (2012). (R)-Cysteate-nitrogen assimilation by *Cupriavidus necator* H16 with excretion of 3-sulfolactate: a patchwork pathway. Arch. Microbiol. 194: 949-957.

19. Merdanovic, M., E. Sauer, and J. Reidl. (2005). Coupling of NAD^+ biosynthesis and nicotinamide ribosyl transport: characterization of NadR ribonucleotide kinase mutants of *Haemophilus influenzae.* J. Bacteriol. 187: 4410-4420.

20. Penfound, T. and J.W. Foster. (1999). NAD-dependent DNA-binding activity of the bifunctional NadR regulator of *Salmonella typhimurium.* J. Bacteriol. 181: 648-655.

21. Reddy V.S. and Saier M.H. Jr. (2012) BioV Suite--a collection of programs for the study of transport protein evolution. FEBS J. 279(11):2036-46. Epub 2012 May 8.

22. Royant, A., P. Nollert, K. Edman, R. Neutze, E.M. Landau, E. Pebay-Peyroula, and J. Navarro. (2001). X-ray structure of sensory rhodopsin II at 2.1-Å resolution. Proc. Natl. Acad. Sci. USA 98: 10131-10136.

23. Russo D., Redman C. and Lee S. (1998) Association of XK and Kell blood group proteins. J Biol Chem. 273(22):13950-6.

24. Saier M.H. and Zhai Y. (2001) A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. *J. Mol. Microbiol. Biotechnol.* 2001 Vol 3 (501-502).

25. Seward, D.J., A.S. Koh, J.L. Boyer, and N. Ballatori. (2003). Functional complementation between a novel mammalian polygenic transport complex and an evolutionarily ancient organic solute transporter, *OSTα-OSTβ.* J. Biol. Chem. 278: 27473-27482.

26. Shlykov, M.A., W.H. Zheng, J.S. Chen, and M.H. Saier, Jr. (2012). Bioinformatic characterization of the 4-Toluene Sulfonate Uptake Permease (TSUP) family of transmembrane proteins. Biochim. Biophys. Acta. 1818: 703-717.

27. Sosso D., Chen L.Q. and Frommer W.B. In: Encyclopedia of Biophysics. Robert G, editor. Vol 5. Springer, 2013. Pp. 2556-2558. Ch.

28. Söding J., Biegert A. and Lupas A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33(Web Server issue):W244-8.

29. Sun M.X., Huang X.Y., Yang J., Guan Y.F. and Yang Z.N. (2013) Arabidopsis RPG1 is important for primexine deposition and functions redundantly with RPG2 for plant fertility at the late reproductive stage. Plant Reprod. 26(2):83-91.

30. Suzuki J., Denning D.P., Imanishi E., Horvitz H.R. and Nagata S. (2013) Xk-related protein 8 and CED-8 promote phosphatidylserine exposure in apoptotic cells. Science. 341(6144):403-6.

31. Thoene, J., R. Lemons, Y. Anikster, J. Mullet, K. Paelicke, C. Lucero, W. Gahl, J. Schneider, S.G. Shu, and T. Campbell. (1999). Mutations of CTNS causing intermediate cystinosis. Mol. Genet. Metabol. 67: 283-293.

32. Thorens B., Guillam M.T., Beermann F., Burcelin R. and Jaquet M. (2000) Transgenic reexpression of GLUT1 or GLUT2 in pancreatic beta cells rescues GLUT2-null mice from early death and restores normal glucose-stimulated insulin secretion. J Biol Chem. 275(31):23751-8.

33. Walker R.H., Jung H.H., Dobson-Stone C., Rampoldi L., Sano A., Tison F. and Danek A. (2007) Neurologic phenotypes associated with acanthocytosis. Neurology. 68(2):92-8.

34. Wang, W., D.J. Seward, L. Li, J.L. Boyer, and N. Ballatori. (2001). Expression cloning of two genes that together mediate organic solute and steroid transport in the liver of a marine vertebrate. Proc. Natl. Acad. Sci. USA 98: 9431-9436.

35. Weinitschke, S., K. Denger, A.M. Cook, and T.H. Smits. (2007). The DUF81 protein TauE in *Cupriavidus necator* H16, a sulfite exporter in the metabolism of C2 sulfonates. Microbiology 153: 3055-3060.

36. Xu Y., Tao Y., Cheung L.S., Fan C., Chen L.Q., Xu S., Perry K., Frommer W.B. and Feng L. (2014) Structures of bacterial homologues of SWEET transporters in two distinct conformations. Nature. 515(7527):448-52.

37. Yee, D.C., M.A. Shlykov, A. Västermark, V.S. Reddy, S. Arora, E.I. Sun, and M.H. Saier, Jr. (2013). The transporter-opsin-G protein-coupled receptor (TOG) superfamily. FEBS J. 280: 5780-5800

38. Zhai Y. and Saier M.H. Jr. (2001) A web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. J Mol Microbiol Biotechnol. 3(2):285-6.