Phylogenetic analysis and database of putative bacterial avidins

Master's Thesis
Degree Programme in Biotechnology
Faculty of Medicine and Life Sciences
University of Tampere, Finland
Tanja Kuusela
October 2017

Acknowledgements

First and foremost, I would like to thank my ever patient and supportive supervisors Vesa Hytönen and Olli Laitinen. Their advice and guidance helped me work through this project that proved longer than anticipated. I thank my family and especially my grandparents for their moral support and encouraging presence. Another, thank you for all my friends who I have met along this journey, your sense of humour and wit has made me laugh, when it has been direly needed. Finally, I would like to thank you, Aapo, for your love, support, and patience when I have entrenched myself at my desk for long hours.

PRO GRADU –TUTKIELMA

Paikka:	TAMPEREEN YLIOPISTO Master's Degree Programme in Biomedical Technology Lääketieteen ja biotieteiden tiedekunta Tampere, Suomi
Tekijä:	Tanja Kuusela
Otsikko:	Phylogenetic analysis and database of new putative bacterial avidins
Sivumäärä:	79 + Liitteet
Ohjaajat:	Apulaisprof. Vesa Hytönen, Dos. Olli Laitinen
Tarkastajat:	Prof. Matti Nykter, FT. Juha Määttä
Aika:	Lokakuu, 2017

Tiivistelmä

Ensimmäinen avidiini havaittiin kanan (*Gallus gallus*) munanvalkuaisesta 1940-luvulla. Tämän löydöksen jälkeen avidiineja on tunnistettu kaikista linnuista, sammakkoeläimistä, matelijoista, kaloista, joistakin selkäjänteettömistä, kahdesta sienilajista, sekä useista bakteereista. Avidiinista on tullut bioteknisten sovellusten perustyökalu, sillä se sitoutuu ligandinsa kanssa ainutlaatuisen tiukasti. Intensiivisestä tutkimuksesta huolimatta tämän proteiinin biologinen tehtävä ei ole kui-tenkaan selvinnyt. Avidiinin toiminnasta antibioottisesti on viitteitä kanalla ja muilla eläimillä, mutta tämän antibioottisen toiminnan mekanismia tai avidiinin merkitystä bakteereilla ei ole pystytty osoittamaan. Tässä tutkielmassa rakennettiin tietokanta oletetuista bakteerisista avidii-neista, fylogeneettinen puu niiden sekvensseistä, ja todennettujen sekä tutkittiin olettettujen avidiini-geenien genomista ympäristöä.

Sekvenssimateriaali kerättiin olemassaolevista tietokannoista. Neljätoista todennettua avidiinisekvenssiä linjattiin rakenteelliseksi MSA:ksi (multiple sequence alignment) T-COFFEE-algoritmin avulla ja MEGA6.0-ohjelman avulla tämän MSA:n pohjalta rakennettiin fylogeneettinen ML(maximum likelihood)-puu. Kattavampi MSA koottiin MUSCLE-algoritmilla ja se sisälsi 113 olettetua avidiinisekvenssiä. Tätä MSA:ta pohjana käyttäen rakennettiin edelleen ML-puu MEGA6.0-ohjelman avulla. Lisäksi 10 genomia, joista oletettuja avidiineja löytyi, valittiin rikastusanalyysiä (engl. enrichment analysis) varten. Avidiinigeenin assosiaatiota GO(Gene Ontology)-termien kanssa tutkittiin Fischerin tarkalla testillä perustuen geeneihin, jotka sijaitsivat enintään 500 bp ylä- tai alavirtaan avidiinigeenistä.

Oletettuja avidiinisekvenssejä löytyi niin lajeista, joista on jo aiemmin raportoitu avidiineja, kuin myös lajeista joista avidiinigeenejä ei ole todennettu. Huomionarvoisia uusia sukuja avidiinigeenin sisältävien bakteerien joukossa olivat *Legionella* ja *Xanthomonas*-suvut. Fylogenia-analyysi osoitti todennettujen avidiinien muodostavat selkeitä alaryhmiä yhdessä useiden aiemmin todentamattomien sekvenssien kanssa. Lisäksi viisi ryhmää muodostui myös todennettujen avidiinien ulkopuolelle, mikä viittaa avidiiniperheen olevan oletettua laajempi. Yllättävästi, sieniavidiinit ryhmittyivät bakteeristen streptavidiinien kanssa muiden eukaryoottien sijaan. Jokaisessa ryhmässä oli lisäksi hyvin eriäviä sekvenssejä. Nämä sekvenssit ja alaryhmät, joissa ei ollut todennettuja avidiinisekvenssejä ovat erityisen kiinnostavia kohteita jatkotutkimukselle. Kahdessa ryhmässä havaittiin laajempia sekvenssimuutoksia β6-säikeessä sekä pitkässä pidennyksessä proteiinin C-päässä. Genomista ympäristöä tarkasteltaessa avidiinigeenejä löytyi niin kromosomeista kuin plasmideista. Rikastusanalyysissä avidiinigeenin kanssa korreloivat mobiilielementteihin ja perusaineenvaihduntaan liittyvät geenit.

MASTER'S THESIS

Place:	UNIVERSITY OF TAMPERE Master's Degree Programme in Biomedical Technology Faculty of Medicine and Life Sciences Tampere, Finland
Author:	Tanja Kuusela
Title:	Phylogenetic analysis and database of new putative bacterial avidins
Pages:	79 + Appendices
Supervisors:	Assoc Prof. Vesa Hytönen, Adj Prof. Olli Laitinen
Reviewers:	Prof. Matti Nykter, PhD. Juha Määttä
Time:	October, 2017

Abstract

Since the discovery of the first avidin in the chicken (*Gallus gallus*) egg-white in the 1940s, avidins have been identified in all avians, amphibians, reptiles, fish, some non-chordate, two fungi, and several bacteria. Avidins have become a staple in biotechnological applications, due to its uniquely tight ligand binding. Despite the intensive study on this protein its biological function remains unknown. Evidence suggests that it could act as an antibiotic agent in chicken and other animals, but the actual mechanism of this function nor its purpose in microbes have not been demonstrated. Here a database of putative bacterial avidins was built, phylogeny of bacterial avidins was constructed, and the genomic context of avidin gene was explored.

The material was collected from the existing databases. A structural MSA (multiple sequence alignment) was constructed with T-COFFEE from a set of 14 verified avidin sequences and a ML (maximum likelihood) phylogenetic tree was built with MEGA6.0 based on this MSA. More expansive MSA was built using MUSCLE from and included a set of 113 putative avidins sequences. A phylogenetic ML tree was constructed from this alignment, again with MEGA6.0. Finally, 10 origin genomes were selected from the set of putative avidins and used in enrichment analysis. The avidin genes' association with GO(Gene Ontology)-terms was tested with the Fischer's exact test based on genes 500 bp upstream and downstream from the avidin gene end.

The putative bacterial avidins sequences included both species with an avidin gene and species avidin has not been reported in yet. Notable groups of new possibly avidin-expressing species included *Legionella* and *Xanthomonas*. The phylogeny showed that the verified avidins form distinct sub groups, each of which contained several previously unreported sequences. Additional five sub groups were present outside of the verified avidins, as well, and this hints the avidin family is more extensive than previously thought. Curious detail was the fungal avidins forming an outgroup of streptavidins, instead of grouping with other eukaryotic avidins. Each sub group contained a few unusually variable sequences that could be of special interest as future research targets. Similarly, the sub groups without functional avidins should be investigated further. Even large scale sequence rearrangements were present in these sub groups: these included complete rearrangement of the β6-strand and long extensions in the C-terminus. Upon inspection of the genomic context, avidin was found in both chromosomes and plasmids. The enrichment analysis brought up avidin genes' correlation with mobile element related genes and housekeeping genes.

Table of contents

Abbreviations

3D	three-dimensional
AA	amino acid
AVD	avidin (*Gallus gallus*)
AVR	avidin related protein (*Gallus gallus*)
BBP-A	biotin-binding protein A (*Gallus gallus*)
BBP-B	biotin-binding protein B (*Gallus gallus*)
Bjavd 1	*Bjavd 1* encoded protein (*Branchiostoma japonicum*)
Bjavd 2	*Bjavd 2* encoded protein (*Branchiostoma japonicum*)
BLAST	basic local alignment search tool
bp	base pair
Bradavd 1	bradavidin 1 (*Bradyrhizobium diazoefficiens*)
Bradavd 2	bradavidin 2 (*Bradyrhizobium diazoefficiens*)
BTSP	bootstrap(ping)
Burkavd 1	burkavidin 1 (*Burkholderia pseudomallei*)
Burkavd 2	burkavidin 2 (*Burkholderia pseudomallei*)
cDNA	complementary DNA
CSC	IT Center for Science
CSS	Cascading Style Sheets
DB	database
DELTA-BLAST	domain enhanced lookup time accelerated BLAST
DNA	deoxyribonucleic acid
FABP	fatty acid binding protein
GUI	graphical user interface
HGT	horizontal gene transfer
Hoefavd	hoefavidin (*Hoeflea phototrophica*)
HTML	Hypertext Markup Language
JSON	JavaScript object notation
JTT	Jones-Taylor-Thornton
K_d	dissociation constant
Lentiavd 1	lentiavidin 1 (*Lentinula edodes*)
Lentiavd 2	lentiavidin 2 (*Lentinula edodes*)
M	molar

MEGA	molecular evolutionary genetics analysis
MPI	metalloproteaseinhibitor
ML	maximum likelihood
MSA	multiple sequence alignment
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
NCBI	the National Center for Biotechnology Information
ND	not determined
NJ	neighbour-joining
NNI	nearest neighbour interchange
PDB	protein data bank
PSI-BLAST	position-specific iterative BLAST
Rhizavd	rhizavidin (*Rhizobium etli*)
Rhodavd	rhodavidin (*Rhodopseudomonas palustris*)
Shwanavd	shwanavidin (*Shewanella denitrificans*)
Streptavd	streptavidin (*Streptomyces avidinii*)
Streptavd v1	streptavidin v1 (*Streptomyces violaceus*)
Streptavd v2	streptavidin v2 (*Streptomyces violaceus*)
Strongavd	strongavidin (*Strongylocentrotus purpuratus*)
Tamavd 1	tamavidin 1 (*Pleurotus cornucopiae*)
Tamavd 2	tamavidin 2 (*Pleurotus cornucopiae*)
Xantavd	xantavidin (*Xanthomonas campestris*)
Xenavd	xenavidin (*Xenopus tropicalis*)
Zebavd	zebavidin (*Danio rerio*)

1 Introduction

Avidins are known for their unrivalled binding of biotin, also known as vitamin B_7 or vitamin H. The first avidin (AVD) was found in the chicken (*Gallus gallus*) egg white and it became a text-book example of tight protein-ligand interaction with its dissociation constant (K_d) that reached remarkably 10^{-15} M (Eakin et al. 1940; Green 1963). After its potential in different applications was realized, other avidins have been sought for and detected across a wide variety of organisms. Avidin-biotin technology is used in purification, detection, and assay technologies in research, but also in diagnostics and pharmaceuticals, only to name few of its practical applications.

The AVD is a small homotetrameric protein with one biotin-binding site at the entrance of each subunit's β-barrel (Livnah et al. 1993). It is considered to belong to the calycin superfamily (Flower 1993). Calycins share the general β-barrel structure and the ability to bind small hydrophobic ligands (Flower 1993; Flower 1996). However, calycins have rather large sequence variation and generally the protein families belonging to this super family share sequence identities of only 20 or 30 % (Flower 1996; Flower et al. 2000). This same trend is seen among the discovered avidins.

The low sequence similarities that fall to the so called "twilight zone" (20–30 %) among avidins make them a relatively difficult targets to identify and align reliably (Chang et al. 2008). Avidins have been discovered many oviparous animals, some non-chordates such as lancelet and sea urchin, fungi, and variety of bacteria including symbionts, marine photosynthetic species, and opportunistic pathogens (Hertz & Sebrell 1942; Nordlund et al. 2005; Venekoski 2009; Sardo et al. 2011; Avraham et al. 2015; Guo et al. 2017). However, high quality phylogenetic tree of the avidin sequences has not been published or proposed.

Interestingly, as studied and utilized as avidin is, its biological function is still unknown. It has been suggested to act as an antibiotic in chicken eggs and lancelet (Tuohimaa et al. 1989; Guo et al. 2017). Avidin gene has also been successfully transferred to plants, such as rice and maize, to act as an intrinsic pesticide (Yoza et al. 2005; Takakura et al. 2012). Furthermore, the amount of available biotin in soil directly correlates with the amount of root eating nematodes and poor growth (Sinkkonen et al. 2014). The avidin secreted by the symbiont bacteria could thus provide a competitive edge for both the host and bacteria as antibiotic and an insecticide. This study aims to explore the avidin family for a clue of its biological function and ease the future study with construction of a new avidin sequence database and identification of potentially interesting targets for future research.

2 Literature review

In the literature review for this study, the avidin protein functional and structural characteristics will be summarized. Furthermore the calycin superfamily, which avidin family is part of, will be briefly introduced. The studies on avidins' biological functions will be outlined. Finally, as the genomic context and the avidin significance for the bacteria will be considered in this study, the specifics of bacterial genome and protein evolution in bacteria will be briefly introduced.

2.1 Avidin

The first avidin (AVD) was isolated from chicken (*Gallus gallus*) egg-white in 1941 after sole egg-white diet was found to cause biotin deficiency in chicks (Eakin et al. 1940). The protein was named avidin (*engl.* avid + biotin) for its unrivalled binding to D-biotin or vitamin B_7, a required vitamin for all living cells, with the dissociation constant (K_d) of approximately 10^{-15} M (Green 1963; Green 1975; Green 1990). AVD is a rather small protein, only 66–69 kDa in size in its native homotetrameric form (Green 1975; Green 1990). Each of its subunits consist of a single 8-stranded β-barrel domain and bind individually one biotin molecule (Livnah et al. 1993; Rosano et al. 1999). AVD's compact structure and remarkably tight-binding to its ligand, as well as existing methods to biotinylate almost any biomolecule has made the avidin both a textbook example of protein-ligand interaction and an important tool in biotechnology. From now on, the abbreviation AVD will be specifically used to refer to the chicken avidin and the name avidin will be used as a generic term for an avidin family member.

The first bacterial avidin, streptavidin (Streptavd), was isolated from *Streptomyces avidinii*, an antibiotic secreting bacteria, in 1964 (Tausig & Wolf 1964; Chaiet & Wolf 1964). Since then, the number of known avidins has increased fast. For all the experimentally verified avidins and citations, refer to the Table 1. Eight avidin family members were identified in the chicken between the 1980s and the early 2000s and further eukaryotic avidins have been found in other avian species, reptiles, amphibians, sea urchin, fish, lancelet and fungi (Hertz & Sebrell 1942; Botte & Granata 1977; Hytönen et al. 2003). Bacterial avidins have arisen in a wide variety of genera including species from symbiotic, marine, and pathogenic niches.

Despite the extensive research on avidin and its wide-spread presence over the kingdoms and environmental niches, the biological function of this protein remains elusive. It has been suggested to have antibiotic qualities, as it renders biotin, unavailable (Tranter & Bourd 1982; Tuohimaa et al. 1989). However, biotin is synthesized by most bacteria, archaea as well as some plants (Streit & Entcheva 2003).

Table 1. The experimentally verified avidins and avidin-related proteins.

	Protein	Origin organism	K_d	Quaternary structure	Source
Eukaryotic	AVD † ‡	*Gallus gallus*	≈ 10^{-15} M	tetramer	(Eakin et al. 1940; Green 1975)
	AVR 1	*Gallus gallus*	≈ 10^{-8} M	tetramer	(Keinänen et al. 1994; Laitinen et al. 2002)
	AVR 2	*Gallus gallus*	≈ 10^{-8} M	tetramer	(Keinänen et al. 1994; Laitinen et al. 2002)
	AVR 3	*Gallus gallus*	<< 10^{-8} M	tetramer	(Keinänen et al. 1994; Laitinen et al. 2002)
	AVR 4/5	*Gallus gallus*	≈ 10^{-14} M	tetramer	(Keinänen et al. 1994; Laitinen et al. 2002)
	AVR 6	*Gallus gallus*	<< 10^{-8} M	tetramer	(Ahlroth et al. 2000; Laitinen et al. 2002; Helppolainen et al. 2008)
	AVR 7	*Gallus gallus*	≈ 10^{-9} M	tetramer	(Ahlroth et al. 2000; Laitinen et al. 2002)
	BBP-A	*Gallus gallus*	ND	tetramer	(Niskanen et al. 2005; Hytönen et al. 2007)
	BBP-B	*Gallus gallus*	ND	tetramer	(Niskanen et al. 2005; Hytönen et al. 2007)
	Xenavd † ‡	*Xenopus tropicalis*	≈ 10^{-13} M	tetramer	(Määttä et al. 2009)
	Zebavd † ‡	*Danio rerio*	≈ 10^{-9} M	tetramer	(Taskinen et al. 2013)
	Strongavd ‡••	*Strongylocentrotus purpuratus*	ND	tetramer	(Veneskoski 2009)
	Tamavd 1 † ‡	*Pleurotus cornucopiae*	ND	tetramer	(Takakura et al. 2009)
	Tamavd 2 ‡	*Pleurotus cornucopiae*	ND	tetramer	(Takakura et al. 2009)
	Lentiavd 1 *	*Lentinula edodes*	ND	ND	(Takakura et al. 2016)
	Lentiavd 2 *	*Lentinula edodes*	ND	ND	(Takakura et al. 2016)
	Bjavd 1 *	*Branchiostoma japonicum*	ND	ND	(Guo et al. 2017)
	Bjavd 2 *	*Branchiostoma japonicum*	ND	ND	(Guo et al. 2017)
Bacterial	Streptavd † ‡	*Streptomyces avidinii*	≈ 10^{-14} M	tetramer	(Tausig & Wolf 1964; Chaiet & Wolf 1964)
	Streptavd v1	*Streptomyces violaceus*	ND	tetramer	(Bayer et al. 1995)
	Streptavd v2	*Streptomyces violaceus*	ND	tetramer	(Bayer et al. 1995)
	Bradavd I † ‡	*Bradyrhizobium diazoefficiens*	≈ 10^{-10} M	tetramer	(Nordlund et al. 2005; Leppiniemi et al. 2012)
	Bradavd II † ‡	*Bradyrhizobium diazoefficiens*	<< 10^{-10} M	mono-/di-/tetramer	(Helppolainen et al. 2008; Leppiniemi et al. 2013)
	Rhizavd † ‡	*Rhizobium etli*	ND	dimer	(Helppolainen et al. 2007; Meir et al. 2009)
	Shwanavd † ‡	*Shewanella denitrificans*	<< 10^{-10} M	dimer	(Meir et al. 2012)
	Hoefavd ‡	*Hoeflea phototrophica*	ND	dimer	(Ahlroth et al. 2000)
	Rhodavd ‡••	*Rhodopseudomonas palustris*	ND	ND	(Sardo et al. 2011)
	Burkavd 1••	*Burkholderia pseudomallei*	ND	tetramer	(Sardo et al. 2011)
	Burkavd 2 ‡	*Burkholderia pseudomallei*	<< 10^{-7} M	tetramer	(Sardo et al. 2011)
	Xantavd ••	*Xanthomonas campestris*	ND	ND	(Helppolainen et al. 2008)

† sequence was used as a query-sequence in the initial BLAST searches

‡ sequence was included in the phylogenetic analysis

* protein was reported after the beginning of the study

•• reported in literature, but not verified experimentally

ND not determined experimentally

For the dissociation constants, only direct measurements with biotin or estimation against were included.

2.1.1 Avidin structure

The 3D-structure of avidin is well preserved across the different organisms, although the amino acid (AA) sequence identities across the family can be as low as 20 % (Helppolainen et al. 2007; Helppolainen et al. 2008). The two most extensively studied proteins in the avidin family, AVD and Streptavd, are both homotetramers with each subunit consisting of a single 8-stranded antiparallel β-barrel domain (Fig. 1, panel A) (Hendrickson et al. 1989; Livnah et al. 1993; Rosano et al. 1999). As the dimer is the asymmetric unit of avidin, the whole structure can be described as a dimer of dimers (Kurzban et al. 1991).

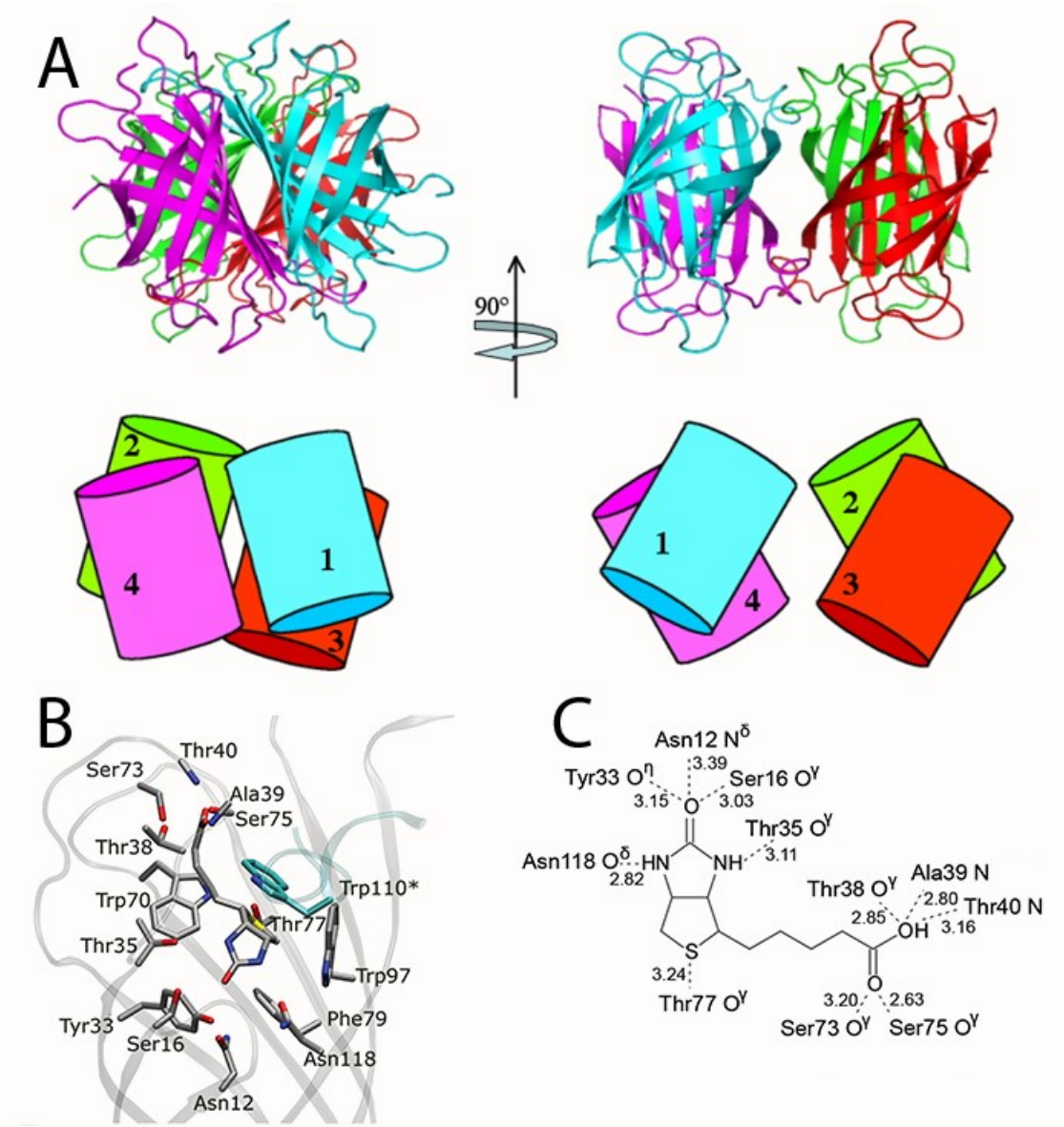


Figure 1. Avidin structure. Each of the four identical avidin subunits are given their own color in the visualization with subunit 1 colored cyan, subunit 2 green, subunit 3 red, and subunit 4 magenta (A) (Eisenberg-Domovich et al. 2005). The subunits 1 and 4 (and similarly 2 and 3) form a structural dimer with many close interactions (A) (Eisenberg-Domovich et al. 2005). The biotin-binding pocket within each subunit β-barrel conforms to the biotin shape closely (B) (Leppiniemi et al. 2011). Several AA residues participate in biotin-binding with direct H-bonds (C) (Eisenberg-Domovich et al. 2005).

The 1/4-interface, between the subunits 1 and 4, forms the structural dimer with extensive interactions between their β-barrel surfaces (Fig. 1, panel A) (Livnah et al. 1993; Rosano et al. 1999). In the AVD, these structural dimers are held together by van der Waals interactions in the 1/3-interface: Met96, Val115, and Ile117; while in the Streptavd the interactions also build upon H-bonds (Table 2) (Livnah et al. 1993; Rosano et al. 1999). Lastly, the 1/2-interface binds together a so-called functional dimer and relies solely on the Trp110 in AVD and the homologous Trp120 in Streptavd (Hendrickson et al. 1989; Livnah et al. 1993; Rosano et al. 1999). The Trp120 resides in the loop 7–8, between the β7 and β8-strands, of the avidin polypeptide. This loop extends to the biotin-binding site in the neighbouring subunit, where the residue participates in stabilization of the biotin bi-cyclic ring (Livnah et al. 1993; Rosano et al. 1999).

The biotin-binding pocket is located at the non-terminus-end of the β-barrel and has to conform to both hydrophobic and hydrophilic regions of the ligand molecule (Fig. 1, panel B) (Livnah et al. 1993; Rosano et al. 1999). The biotin main bulk consists of a dual ring with ureido and tiophene moieties (https://pubchem.ncbi.nlm.nih.gov/compound/biotin, 13.3.2017) (Kim et

Table 2. The key amino acids participating in structure and biotin-binding functions of avidin and streptavidin.

Amino acid role	AVD	Streptavd
Interface 1/3	Met96	Gln107
	Val 115	Val125
	Ile117	His127
Interface 1/2	Trp120	Trp110
H-bond network	Trp10	Trp21
	Asn12 *	Asn23 *
	Asp13	Gln24
	Ser16 *	Ser27 *
	Tyr33 *	Tyr43 *
	Thr35 *	Ser45 *
	Trp70	Trp79
	Ser75	Ser88
	Thr77	Thr90
	Asn118 *	Asp128 *
Van der Waals interactions with ligand	Leu14	Leu25
	Trp70	Trp79
	Phe79	Trp92
	Trp97	Trp108
	Leu99	Leu110
	Trp110	Trp120

* participates in biotin-binding through H-bonds

al. 2016). The ring structure is hydrophilic and rigid, while the valeryl tail brings both flexibility and hydrophobicity to the molecule before the final hydrophilic carboxyl group at its end.

In the absence of the ligand, the binding pocket is occupied by five water molecules that escape through a water channel at end of the barrel upon biotin arrival (Rosano et al. 1999; Hyre et al. 2002). The biotin is prevented from leaving through this same end by a vast network of H-bonds (Rosano et al. 1999). Albeit several of the individual AAs participating in the network vary between the verified avidins, the conserved Trp21 (in Streptavd and Trp10 in AVD) not only contributes to the sealing H-bond network, but also forms a bulky steric obstacle at the non-open end of the barrel (Livnah et al. 1993; Rosano et al. 1999). Furthermore, the steric obstacle together with the H-bonds prevent the solvent from re-entering the binding site with bound biotin (Rosano et al. 1999).

The bi-cyclic ureido-tiophene moiety of biotin resides at the very bottom of the sealed barrel structure (Livnah et al. 1993). The pocket around the biotin is stabilized by the same H-bond network that seals the end of the barrel (Livnah et al. 1993; Rosano et al. 1999). The residues Asn23, Ser27, Tyr43, Ser45 and Asp128 are especially important in the Streptavd, as they also form direct H-bonds with the biotin ring structure (Fig. 1, panel C) (marked with an asterisk in Table 2) (Rosano et al. 1999).

The hydrophobic van der Waals interactions in the binding pocket's narrow region also help to stabilize the bound biotin (Livnah et al. 1993; Rosano et al. 1999). The most important residues participating in these interactions are Leu25, Trp79, Trp92, Trp108, Leu110, and Trp120 in the Streptavd (Rosano et al. 1999). The four Trp residues sandwich the bi-cyclic ring for a ring-stacking effect (Rosano et al. 1999). The 120 (or Trp110 in the AVD) is offered by the neighbouring subunit's loop 7–8 (Livnah et al. 1993; Rosano et al. 1999). Lastly, the loop 3–4 region at the non-terminal end of the β-barrel is a highly flexible structure that enables the biotin entrance to the binding pocket with ease, but upon biotin arrival the loop stabilizes and shields biotin from the solvent (Livnah et al. 1993; Rosano et al. 1999).

2.1.2 Avidin protein family

Together with lipocalins, metalloproteaseinhibitors (MPIs) and fatty acid binding proteins (FABPs) the avidin family belongs to the structural superfamily of calycins (Flower 1993). The common denominator for this superfamily's 3D-structure is an antiparallel β-barrel with 8 (or 10 in the FABPs) β-strands (Flower 1993; Flower 1996). Most of the calycins bind small hydrophobic molecules deep in the barrel structure. While avidins show similar structure, the avidin β-barrel has a more round cross section and significantly shorter loop 1–2 (Fig. 2)

(Flower 1993). Lipocalins also contain an elongation with an alpha-helix at the C-terminus of the barrel, another feature that avidins lack (Fig. 2) (Flower 1996). The sequence identity in each calycin family is very low, reaching even 12 to 20 % among lipocalins (Flower 1996; Flower et al. 2000). As mentioned earlier, the identities among avidins are in line with this falling even as low as 20 %. Due to these low sequence identities, avidins and other calycins are difficult targets to identify and distinguish solely by sequence comparisons and alignments.

Avidin family contains currently 30 experimentally verified members from 20 different species (Table 1). However, there has been no published attempts to organize them phylogenetically aside the chicken avidins (Wallén et al. 1995; Ahlroth et al. 2001; Hytönen et al. 2005). In a master's thesis work by Tiwari in 2015, a more comprehensive phylogenetic tree was constructed. A rough division between the eukaryotic and the bacterial avidins is seen in this analysis. Although, the fungal tamavidins (Tamavd 1 and Tamavd 2) cluster together with the bacterial streptavidins (Streptavd, Streptavd v1, and Streptavd v2). The dimeric avidins, so far encountered only in bacteria, roughly form their own group, as well.

Aside the avidins themselves, the avidin family is considered to include fibropellins as described by Yanai et al. in 2005. The fibropellins are multidomain proteins found from the sea urchin (*Strongylocentrotus purpuratus*) with a C-terminal avidin-like β-barrel structure (Bisgrove et al. 1995; Yanai et al. 2005). The protein is involved in the formation of the apical lamina around the sea urchin embryo (Bisgrove et al. 1995). Interestingly, the fibropellins' avidin-like domain does not bind biotin, but when it is expressed in itself it forms homotetramers (Yanai et al. 2005). The tetramerization is expected to occur also in the full sequence fibropellins and thus the domain could work as a structural component. Yanai et al. further suggest this is a molecular case of exaptation, where an existing physical feature is modified to adapt for a new function.

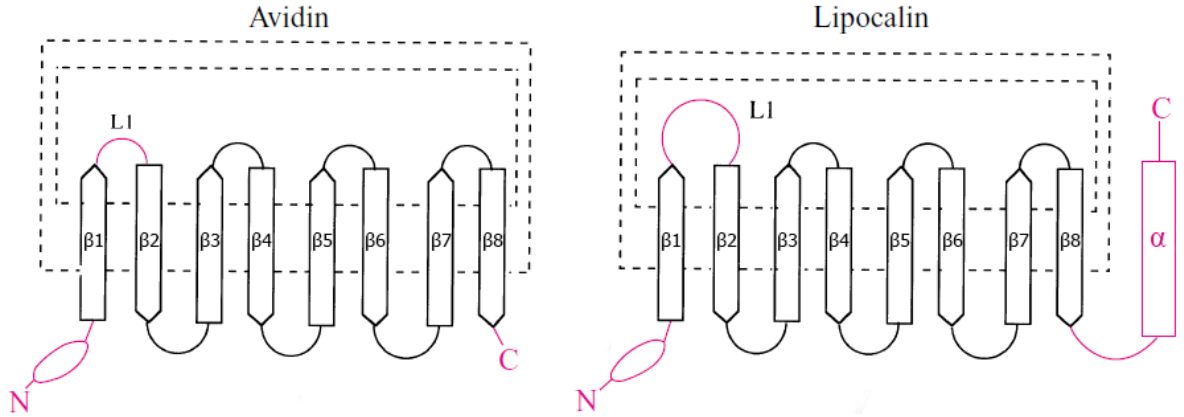


Figure 2. Comparison of avidin and lipocalin protein topology diagrams. Avidins and lipocalins show similar topology in their 8-stranded β-barrel structure. Figure adapted from Flower's review from 1996. Segments that are different between different calycin families, N-terminus, loop 1–2, and C-terminus, are with marked magenta.

2.1.3 Avidin biological function

In oviparous animals, avidin has been suggested to act as an antimicrobial agent that protects the eggs from bacteria within the direct environment (Tuohimaa et al. 1989). Although, no clear mechanism for the anti-microbial function has been demonstrated, avidin has been assumed to render the essential vitamin, biotin, unavailable. The proposition is supported by evidence that chicken, oviductal tissue especially, produces avidin in response to bacterial, viral, and environmental stress, as well as progesterone production (Korpela et al. 1981; Korpela et al. 1982; Tuohimaa et al. 1989; Kunnas et al. 1993). Furthermore, the egg laying order correlates with the survivability of the offspring and the avidin content (Wet & Hsu 1970) . Specifically the first laid egg has higher avidin content as well as a better survival rate.

Further evidence to support avidin's probable antimicrobial role in the animal kingdom was discovered recently, as a new avidin family members Bjavidin 1 and Bjavidin 2 (Bjavd 1 and Bjavd 2, respectively) were discovered from a lancelet species *Branchiostoma japonicum* (Guo et al. 2017). In this species, both *Bjavd 1* and *2* genes were found to be expressed in response to bacterial and heat shock stress. Interestingly, the protein appeared to also recruit macrophages to the site of infection and thus acted as an opsonin in this lancelet.

While no avidin gene has yet been found in any plant species, transgenic avidin-expressing crops, including maize and rice, have been constructed (Hood et al. 1997; Yoza et al. 2005; Martin et al. 2010; Takakura et al. 2012). In both, the avidin made the crop resistant to storage insect pests (Yoza et al. 2005; Takakura et al. 2012). The significance of transgenic avidin in agricultural pest control is further reviewed by Martin et al. in 2010. Interestingly, bacterial avidins of a symbiotic species may, in nature, participate in protecting the plant host. In an ecological study comparing two legumes, an invasive alien species *Lupinus polyphyllus* and the native *Lotus corniculatus*, a correlation between biotin availability and root feeding nematodes was found with the invasive *L. polyphyllus* having less biotin available in its rhizosphere (Sinkkonen et al. 2014). Furthermore, avidin genes has been found in two of the legume associated root nodular symbiotic bacterial species, *Bradyrhizobium japonicum* and *Rhizobium etli* (Nordlund et al. 2005; Helppolainen et al. 2007; Helppolainen et al. 2008). Hypothetically, the avidin could render biotin unavailable in the rhizosphere and thus help control the population of the pest nematodes. Alternatively it could accumulate in the root nodules making them toxic upon consumption. However, the mechanism of function has not been verified and does not explain the avidin presence in the marine nor the pathogenic bacteria.

In fungi, the tamavidins (Tamavd 1 and Tamavd 2), discovered from the edible mushroom *Pleurotus cornucopiae*, have been suggested to protect from phytopathogenic fungi (Takakura et al. 2009; Takakura et al. 2012). These two fungal avidins were later proven to work as pesticides against common fungal insect antagonists, as well (Bleuler-Martinez et al. 2012). The authors thus proposed the avidins might serve as protective agents in this kingdom as well.

2.2 Protein evolution

The basic structural or functional building block of protein is a domain and it is a crucial unit also in the concept of protein evolution. As, domains are characterized as conserved segments of protein that form a distinct fold with a level of functional and structural autonomy, they also can evolve independently. The protein evolution is driven by the cumulative mutations in the encoding genes leading to amino acid substitutions, deletions, and insertions. This eventually causes changes in the protein or domain structure or function. The existing protein domains can be classified to domain families and many of these are found across the kingdoms (Ponting & Russell 2002). This suggests the widely spread protein domains are of ancient origin and either very adaptable or crucial for the cells (Ponting & Russell 2002; Vogel et al. 2004).

However, the development of new functions is not limited to the rise of new domains. Genomic rearrangement events, such as recombination and gene duplication, can bring multiple domains together to create new domain combinations significantly speeding up the protein evolution (Vogel et al. 2004). Duplication events can also speed the rate of allowed mutations, since original function can be performed with one copy while the other can undergo rapid changes (Pál et al. 2006). Similarly, speciation is important in the protein evolution, as it leads the protein copies to adapt to the selective pressures of different environments. The duplication events lead to the emergence of ortologues, or the homologues within a species, while the speciation is the usual source of paralogues, or the homologues across species. In the modern view the paralogues can also originate from the horizontal gene transfer (HGT) as reviewed by Koonin et al. in 2002 (Koonin et al. 2002). While HGT has long been accepted to occur frequently between prokaryotes, it has now been demonstrated to happen between eukaryotes, fungi, plants and even more recently across kingdoms (Richardson & Palmer 2007; Keeling & Palmer 2008; Dunning Hotopp 2011; Xi et al. 2012; Gao et al. 2014; Qiu et al. 2016).

One of the most utilized methods to uncover the underlying duplication and speciation events within a protein family is sequence level phylogeny. It can be used to assess sequence relatedness, identify paralogues and ortologues, and when compared with taxonomical

phylogeny the HGT and gene duplication events can be considered. As the evolutionary distance between sequences grows, the multiple sequence alignments (MSA) have diminishing certainty, since the sequence identities fall. This further impacts the reliability of the phylogenetic analyses, as they rely on high quality MSAs. When the sequence identities drop between 20 and 30 %, into a so-called "twilight-zone", the sequence alignment must be improved with structural information (Chang et al. 2008).

In this context, the avidins are characteristically a protein family that consists of a single domain with the exception of fibropellins. They have been found in a variety of environmental niches and across the kingdoms: in bacteria, oviparous animals, lancelets, and even fungi. The fibropellins that contain the avidin-like structural domain show the family is also functionally diverse. All of this together with the low sequence identities suggest that avidins are evolutionally rather old. While avidins have been found in so many different species, it is also notable that no avidin has yet been discovered from plants, archaea, nor mammals. The low sequence identities and the uneven pattern of presence make the avidin family especially difficult target for accurate phylogenetic analysis and interpretation.

2.2.1 Bacterial genome in protein evolution

The bacterial genome is considered to consist of individually replicating entities called replicons that can be divided in chromosomes and plasmids. The chromosomes replicate intrinsically with the cell's own mechanisms and contain most of the translation and house-keeping genes (Rocha 2004; Rocha 2008). While most of the bacteria contain only one chromosome, secondary chromosomes are also present in other species. These secondary chromosomes often contain fewer house-keeping genes and more adaptive genes (Casjens 1998; Rocha 2004; Rocha 2008). Similarly, plasmids often contain competitive or adaptive gene repertoire (Rocha 2004; Rocha 2008; Smillie et al. 2010).

The gene organization within the replicons is further optimized for function in bacteria (Rocha 2004; Rocha 2008; Darmon & Leach 2014). The essential genes that require constant expression reside near the replication origin for enhanced expression rates and accessibility (Rocha 2008). While the genome is undergoing replication, these genes will exist as multiple copies that can be expressed simultaneously. Meanwhile, the situational genes tend to shift away from the origin (Rocha 2008). Additionally, the bacterial genes of related function or response tend to cluster together into operons that are expressed concurrently (Rocha 2008; Darmon & Leach 2014). A conserved gene organization between different species is specifically called synteny. In bacteria, the synteny outside of the operons is usually rapidly lost

and thus preserved gene organization signals selective pressure for the genes to be expressed together (Darmon & Leach 2014).

The rearrangement and evolution of the bacterial genome often is directed by the environment of the species as reviewed by Liò in 2001. In a restricted, isolated, and stable niche the genome tends to shrink and lose genetic material, first through the deleterious mutations and later by complete gene loss. At the same time, the bacterial species is also shielded from the HGT and acquisition of new genetic material. Intracellular parasitic or symbiotic relationships are good examples of these kind of environments and correlate with shrinking and more compact genomes. On the other end of the scale, are the bacteria in both highly diverse and contested environments, such as rhizosphere with abundance of competing species and microenvironments such as mycelium, soil and plant root nodules. The species in these interface environments often experience expanding of the genome and accumulation of both secondary chromosomes and plasmids with adaptive genes.

Interestingly, at least two of the verified avidin-expressing species *Rhizobium etli* and *Burkholderia pseudomallei*, were mentioned in literature as example cases of bacterial species with large and adaptive genomes (Liò 2002). These species are all living in rather dynamic and varied environments: the *R. etli* resides in the rhizosphere and *B. pseudomallei* is a soil bacteria also capable of causing systemic infection in humans (Liò 2002). *B. pseudomallei* contains 2 to 3 chromosomes, while *R. etli* contains one primary and one secondary chromosome and additionally a plasmid (Liò 2002).

3 Objectives

The aims of this study included:

1. Identification new putative bacterial avidins
2. Construction of avidin-specific sequence database
3. Construction of a comprehensive high quality MSA of verified avidin sequences
4. Construction of a phylogenetic tree of the putative sequences
5. Inspection of avidin gene's genomic context in bacteria

On broader scale, the database, phylogeny, and genomic context were studied to give a clue about the biological function of avidin protein. Similarly, the avidin database and the collection of new putative avidin sequences is aimed to ease identification of new targets for research and verification.

4 Materials and methods

A set of sequences for verified avidins was collected as the starting material based on a literature sweep. This set included six eukaryotic avidin sequences, AVD (P02701, UniProtKB); Xenavd (A7YYL1, UniProtKB); Zebavd (E7F650, UniProtKB); Strongavd (not published); Tamavd 1 (B9A0T6, UniProtKB); and Tamavd 2 (B9A0T7, UniProtKB), as well as eight bacterial avidin sequences, Streptavd (P22629, UniProtKB); Bradavd 1 (Q89IH6, UniProtKB); Bradavd 2 (Q89U61, UniProtKB); Rhodavd (Q218I6; UniProtKB); Rhizavd (Q8KKW2, UniProtKB); Shwanavd (Q12QS6, UniProtKB); Hoefavd (A9D857, UniProtKB); and Burkavd 1 (Q3JRB6, UniProtKB) (The UniProt Consortium 2017). While more avidins have been reported and even experimentally verified, this set was selected to focus on bacterial sequences with only few rather dissimilar eukaryotic avidins. Strepavd v1, as well as Streptavd v2, were excluded since their sequence is highly similar to that of Streptavd.

The sequence set of the verified avidins was used to query further putative avidin sequences that were used to build a new avidin specific database, DATAvidin. Later a set of putative avidins was selected from the DATAvidin contents for further analysis. The verified avidins set and the putative avidins set were used to construct a structural MSA and non-structural MSA respectively and the further phylogenetic cladogram trees. Nine available full genomes among the putative or verified avidin containing species were selected to assess the genomic context of the avidin genes further. Specifics of each method and analysis is specified further below.

4.1 BLAST-queries

Nine verified avidin sequences, streptavidin (P22629, UniProtKB); bradavidin I (Q89IH6, UniProtKB); bradavidin II (Q89U61, UniProtKB); rhizavidin (Q8KKW2, UniProtKB); shwanavidin (Q12QS6, UniProtKB); avidin (P02701, UniProtKB); zebavidin (E7F650, UniProtKB); xenavidin (A7YYL1, UniProtKB); and tamavidin 1 (B9A0T6, UniProtKB), were used as the query sequences (The UniProt Consortium 2017). For protein queries, the domain enhanced lookup time accelerated basic local alignment search tool (DELTA-BLAST) algorithm was chosen, as it takes into account protein domain features and is thus beneficial in searching for related sequences with low identity (Altschul et al. 1990; Boratyn et al. 2012). All non-redundant protein databases were used as a search set and this included RefSeq, Protein Data Bank (PDB), GenBank, and UniProtKB (Berman et al. 2000; Benson et al. 2005; O'Leary et al. 2016; The UniProt Consortium 2017). The search was limited to bacteria and the

maximum target sequence limit was set to 5000, BLOSUM62 was used as the scoring matrix, and the rest of the parameters were set to automatically adjust for short input sequence. The query was further refined four times with PSI-BLAST algorithm, with E-value cut-off of 0.01 and required identity greater than 19 % (Przybylski & Rost 2008). Sequences annotated as pepsins were excluded between iterations to improve the accuracy, since the pepsin annotated hits were poor matches, long proteins and very dissimilar to avidins. The final results were saved into fasta-formatted files with same cut-off values as between PSI-BLAST iterations. Nucleotide sequences were searched for with tBLASTn algorithm against all non-redundant databases again limited to bacteria and the databases included Genbank, The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database, and DNA Data Bank of Japan (DDBJ) (Benson et al. 2005; Kanz et al. 2005; Mashima et al. 2017). Same parameters were used as with protein queries.

4.2 Sequence processing

The BLAST results were further processed with Python (3.4) language's Biopython package. First the result sets were filtered to remove duplicates between them and the rest were combined into one fasta-file for both protein and nucleotide sequences. Any sequences corresponding to synthetic proteins or modified organisms were removed, as they were not the focus of this study. Special care was seen in searching for correspondence between the nucleotide and protein sequences. The nucleotide sequences that represented partial duplicates of a protein sequence were removed to prevent further duplicates in the next step. All protein sequences were inspected to retrieve the original genomic features and their full nucleotide sequences. Similarly, the genomic position for each nucleotide sequence was sought and the sequences were replaced with a previously annotated full complementary deoxyribonucleic acid (cDNA) feature, if such was present. The nucleotide sequences shorter than 300 bp (corresponds to 100 AA of protein sequence) were also extended from the genomic context, when open reading frame allowed. Lastly the nucleotide sequences that did not yet have a corresponding protein sequence were translated and these translated sequences were added to the protein set. In the end, GenBank-files were sought for all the fasta-formatted sequences.

4.3 DATAvidin database

MariaDB (v. 10.0.20) was chosen as the language for the DATAvidin database, since it is an open source fork of the popular MySQL. The database was further built with SQL commands and HeidiSQL graphical user interface (GUI) for MariaDB. Altogether, three tables were

created in the database: prot_seqs (storing the protein sequences), nucl_seqs (storing the nucleotide sequences), and seq_feats (storing the sequence features). Biopython was used to read the information from the sequence data set GenBank-files and format the information in SQL. The SQL command files were then used to transfer the digested information to DATAvidin database. Feature information, on the other hand, was collected by hand from articles and transferred manually to DATAvidin via HeidiSQL interface.

The GUI for the DATAvidin was built in django (v. 1.9.2), a web-development package for Python (3.4); Hypertext Markup Language (HTML); Cascading Style Sheets (CSS) language; javascript with its jquery package; and JavaScript Object Notation (JSON). Django performed the interaction between the database and GUI while HTML and CSS were the main building blocks of the interface itself. Lastly, javascript and jquery handled the interactive functions of the GUI, while JSON served the information to the interactive interface.

External programs were used to perform the local BLAST searches and the alignments in the GUI of DATAvidin. The standalone version of BLAST, BLAST+, is used to perform BLAST queries of the DATAvidin sequences (Camacho et al. 2009). Meanwhile, the alignments utilize ClustalO (v. 1.2.4). Lastly, CSC – IT Center for Science Pouta server was used to host the database.

4.4 Multiple sequence alignment

Two multiple sequence alignments were constructed from the two different sequence sets. Structural MSA used the set of verified avidins, while a more comprehensive MSA was built upon the larger set of the putative avidins identified in this study. The set of putative avidin sequences was refined iteratively by visual inspection after aligning the full set with MUltiple Sequence Comparison by Log-Expectation (MUSCLE) (Edgar 2004). Refinement included removal of too short or too similar sequences as well as too variant sequences that caused difficulties in the alignment process. After the refinement, the actual MSA was constructed based on the set.

4.4.1 Structural alignment

The structural MSA was constructed from the verified avidins set of 14 sequences with T-Coffee (Notredame et al. 2000). T-Coffee was run in Expresso-mode that uses the 3DCoffee and several other alignment algorithms to incorporate both structural and sequence alignment information to the final MSA (Armougom et al. 2006). The BLAST query was set to local, so no new sequences were incorporated outside the given set of verified avidins and the search for

matching Protein Data Bank (PDB) structures was set to automated mode. The structures fetched and used by Expresso in the structural MSA construction were 1vyo for AVD, 4dne for Streptavd, 2y32 for Bradavd I and Rhodavd, 4ggz for Bradavd II, 3ew2 for Rhizavd, 3szj for Shwanavd, 4z6j for Hoefavd, 2uz2 for Xenavd, 4bj8 for Zebavd, 2fhl for Strongavd, 2szc for Tamavd 1 and Tamavd 2. No good enough match for Burkavd was found from PDB structures. All available algorithms suited for AA sequence alignment were used to improve the pairwise alignment quality: sap_pair, TMalign_pair, mustang_pair, mafft_msa, clustalW_msa, muscle_msa, t_coffee_msa, bestpair4prot, and clustalW_pair. Maximum length for the alignment was set to default 2500 and the slow evaluation mode was used to improve the alignment quality further. Finally the MSA was cleaned up manually with AliView (Larsson 2014). The cleaning consisted of removing gaps from the unaligned N- and C-terminal ends in the sequences.

4.4.2 Alignment of the new putative avidins against the structural alignment

The alignment of the putative avidin sequences was constructed using the structural MSA of verified avidins as seed alignment. MUSCLE was again used to align each sequence in the putative avidins set against the profile of verified avidins. Sequence type was set as protein, maximum iterations limited to 1000 and maximum trees to 100. After each alignment, the MSA was inspected visually in AliView. The gaps in sequence ends were removed and if needed the position of biotin-binding and conserved AA homologues were used to adjust the MSA.

4.4.3 Visualization

The alignments were visualized using Jalview 2 to highlight the more conserved positions. The MSA was coloured by the conservation with 30 % cut-off threshold using the BLOSUM62 matrix for similarity. The sub sets of the full putative avidins MSA were visualized separately to better show the group specific features, as sequence variability was too high in the full set. The sub sets were chosen according to phylogenetic sub groups (see below) and visual inspection. Finally, the alignments were further visually processed in InkScape.

4.5 Phylogenetic analysis

Phylogenetic analysis was performed for both the verified avidins set and the putative avidins set in MEGA6.0 using the structural and full MSA, respectively (Tamura et al. 2013). For both trees the maximum likelihood (ML) algorithm was used as well as same parameters: Jones-Taylor-Thornton (JTT) model adjusted for site-specific AA sequences was used as the

substitution model, phylogeny quality was tested with bootstrapping (BTSP) with 1000 replications, rates among sites were set gamma distributed with invariant sites, gaps or missing data was handled with partial deletion while site coverage cut-off was set to 95 %, branch swap filter was strong, and ML heuristic method used the Nearest-Neighbour-Interchange (NNI) with initial tree calculated with the default neighbour-joining (NJ) method (Waterman et al. 1976; Saitou & Nei 1987; Jones et al. 1992; Sullivan 2005; Pearson 2013). The trees were transformed in the end to cladograms, as the evolutionary distances would not be descriptive with the low identities among the sequence sets. Two sub branches of putative avidins with sequence extensions were set as the outgroup of the tree.

4.6 Enrichment analysis

The set of genomes that were chosen to be assessed in enrichment were picked across the different sub branches present in the phylogenetic cladogram trees constructed (see above). The organisms chosen were *Bradyrhizobium diazoefficiens* (BA000040, GenBank), *Ralstonia eutropha* (CP000090–93, GenBank), *Rhizobium etli* (CP001074–77, GenBank), *Methylobacterium extorquens* (CP001298–1300, GenBank), *Catenulispora acidiphila* (CP001700, GenBank), *M. mediterranea* (CP002583, GenBank), *Ralstonia pickettii* (CP00667–69, GenBank), *Legionella pneumophila* (CR628336–38, GenBank), and *Xanthomonas fuscans* (FO681494–97, GenBank) (Benson et al. 2005). The genomic features from these organisms and their assemblies were pooled all together and separately from the present avidin (putative or verified) gene's vicinity. The vicinity of the avidin gene was defined as 500 bp upstream from gene beginning and 500 bp downstream from the gene end. Gene Ontology or GO-terms were searched for each feature. If the feature was not annotated to any GO-terms, the annotations for PFAM, IPR, or TGRFAM terms were sought (Ashburner et al. 2000; Haft et al. 2003; Finn et al. 2016; Finn et al. 2017). These were then mapped to corresponding GO-terms. Fischer's exact test was performed to evaluate, if features annotated to a certain GO-term cluster significantly more often with avidin than could be expected by random distribution. Biopython was used processing and analysing the data.

5 Results

The results of this study included a new database of avidin sequences collected from publically available databases, MSAs of the avidin sequences, and phylogenetic trees based on these. The trees were used to identify sub groups within avidin family. Finally an enrichment analysis was performed to assess avidin genes' association with genes of certain cellular functions.

5.1 BLAST queries

Queries were run against both protein and nucleotide databases with a set of nine verified avidins sequences. For the protein queries the amount of hits varied between 285 and 303, while for the nucleotide queries the amount of hits varied between 13 and 182. As the pooled query results contained a high amount of redundancy, the previously collected protein and nucleotide sequences were processed to obtain a cleaned up set of unique 213 nucleotide and 946 protein sequences. This data together with the set of verified avidin sequences was used as the material for both building of the DATAvidin-database and later analyses.

5.2 DATAvidin: New database on bacterial avidins

The DATAvidin database was constructed from the putative avidins sequence set. Altogether, there was 415 protein entries that were identical, full or partial duplicates of another sequence entry in the database. Sequences in the database originated from 369 bacterial strains. The full sequence pool was six times the size of the strain pool and almost 12 times the size of the species pool. All of the 213 nucleotide entries are present also as protein sequences. The database contents and sources are roughly described in the Figure 3.

In addition, 47 sequence feature entries were collected into the database. The features include β-strands, biotin-binding amino acid residues in AVD and Streptavd, and positions that are generally conserved among the experimentally verified avidins. The feature data table contains a flexible comment segment to record information of mutations at the feature position that may affect the avidin's physicochemical properties.

The graphical user interface to the database can be accessed in the IP-address: http://86.50.169.79:3000. Simple workflow suggestion for the database is described in the Figure 4. The graphical user interface (GUI) contains functionality for BLAST queries and alignments with you own sequences of interest or the DATAvidin protein sequences. The web GUI and its features are described in the Figures Appendix B1 – B6.

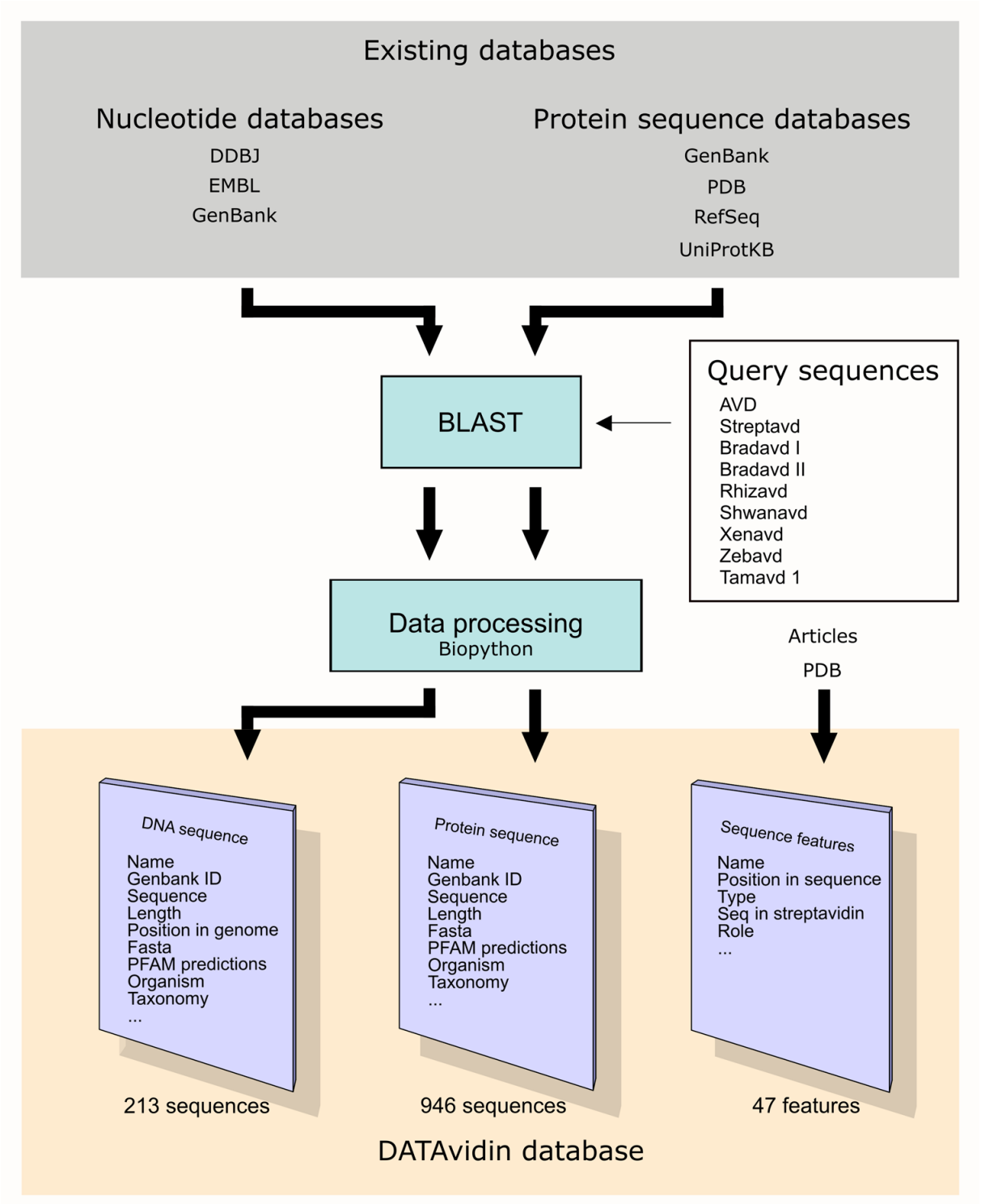


Figure 3. DATAvidin-database contents.
Source material for the DATAvidin database was collected through DELTA-BLAST queries with nine different query sequences of experimentally verified avidins. Queries were run from the publicly available databases. The data obtained from the query results, including fasta- and genbank-files, were further processed to remove any clearly redundant or artificial sequences, to complete any partial sequences from matching genomes, and to link protein and DNA entries to each other. Additionally, the sequence feature information was manually collected from articles and PDB-files.

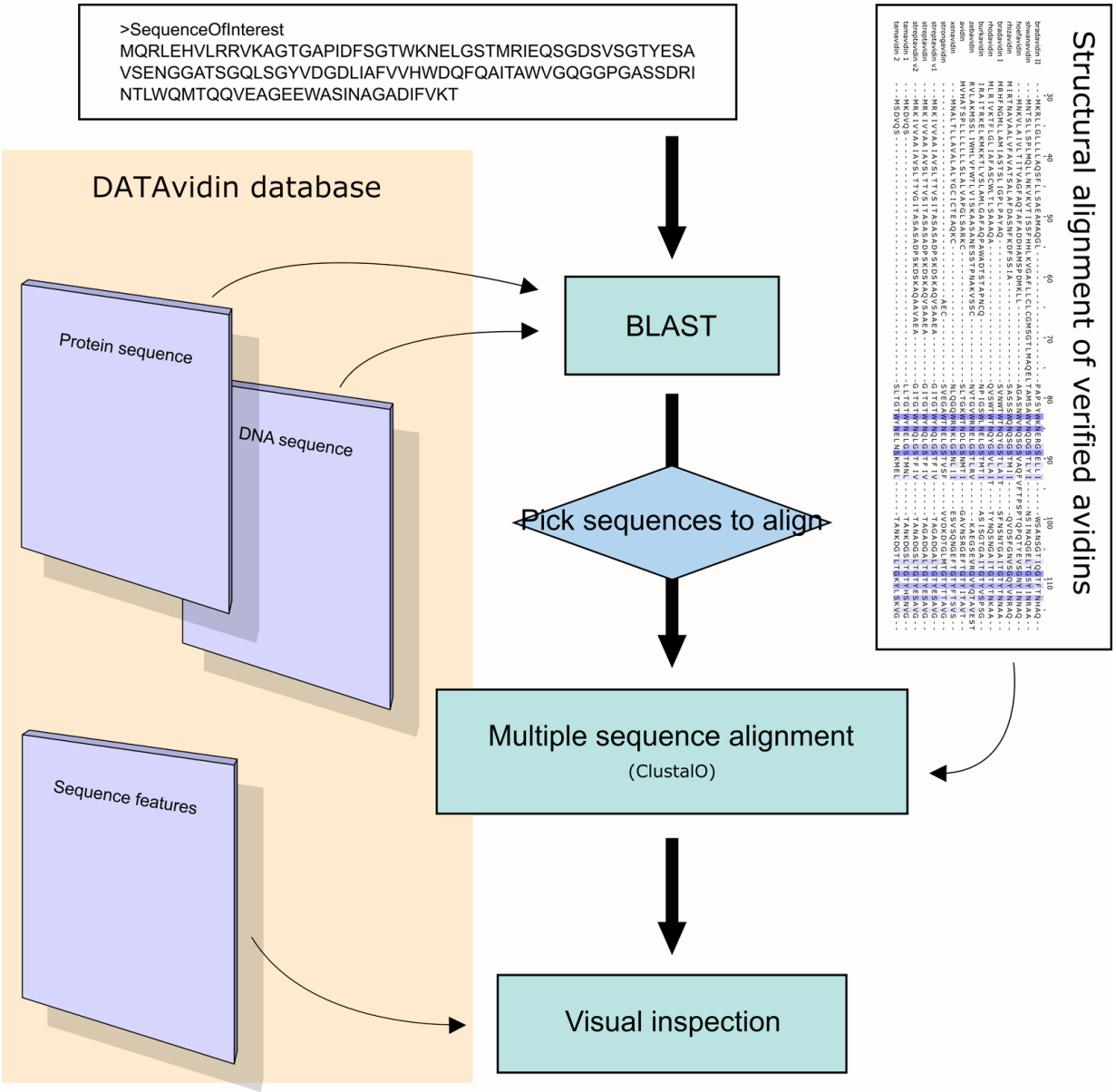


Figure 4. Suggested workflow when using the DATAvidin web-interface.
A sequence of interest can be used in fasta-format to query the database with a local BLAST query or to be directly aligned against the structural multiple sequence alignment (MSA) profile of the experimentally verified avidins. Further interesting sequences can be chosen from BLAST results and aligned against the structural MSA profile. Finally the site provides a simple visualization of the the alignment for visual inspection of the features in avidin sequences aligned with the sequence of interest.

5.3 Phylogenetic and MSA results

A structural multiple sequence alignment (MSA) of the verified avidins sequence set was used to construct a phylogenetic cladogram tree (Fig. 5, Fig. 6). The bootstrap (BTSP) values in this tree were between 36 and 100. 73 % of the nodes had BTSP values greater than or equal to 70 % and 18 % of them had values drop below 50 % (Table 3). A more comprehensive tree was constructed from a MSA (data not shown) that included also the putative avidins sequence set (Fig. 7 and Fig. 8).

BTSP values for this tree were between 7 and 100 with 50 % of the values greater than or equal to 70 (Fig. 8) (Table 3). Only 31 % of the nodes had BTSP values lower than 50 %. Both trees showed comparable topology (Fig. 6, Fig. 8). Strongavidin (Strongavd) was the only leaf node with inconsistent location between the two trees. However, even Strongavd remained within the same clade, which consisted of animal, fungal and *Streptomyces* derived avidins.

The sequences in the phylogenetic trees are named by the origin organism by taking the first four letters from the genus and the species names. In the case of experimentally verified avidins, the name previously mentioned in the literature is used. When more sequences originated from the same organism and sequencing assembly, the sequences are separated with a number at the end of the name. Notably, the numbering starts from 0.

Table 3. The Bootstrap values within and between the branches in the phylogenetic cladograms.
The BTSP (bootstrap) values in the table are expressed as percentages of supporting trees out of all replications. For each set of nodes, the minimum and maximum BTSP values are given, as well as, percentages of BTSP values above 70 % and below or equal to 50 %. The sets of nodes include the branches considered as avidin sub groups and the full trees of both putative and verified avidins. Additionally the nodes between the sub group branches is included as a set for comparison.

	Percentage of nodes with bootstrap values		**Bootstrap values**	
Branch/Set of nodes	**> 70**	**≤ 50**	**Max**	**Min**
Dimeric avidins	53	27	100	33
Bradavidins 1	50	33	98	39
Burkavidins 2	52	35	99	39
Fungal and streptavidins	61	23	100	35
Animal avidins	33	33	99	44
Burkavidins 1	52	32	99	22
Metavidins	100	0	99	97
Legavidins	70	10	97	42
Bradavidins 3	53	20	99	22
β6 avidins	33	0	93	52
Extended avidins	23	53	99	23
Between branches	30	50	95	7
Full tree (putative avidins set)	50	31	100	7
Verified avidins tree	73	18	100	36

5.3.1 Avidin clades and bacterial species

Eleven distinct clades could be identified from the tree constructed from the putative avidins sequence set. Six included at least one previously verified avidin and were clustered together in a superclade with **burkavidins 1** clade, which did not contain a verified avidin. These verified avidin clades were named, clockwise, **dimeric avidins** (gold), **bradavidins 1** (brown), **burkavidins 2** (orange), **fungal and streptavidins** (red), **animal avidins** (pink), and **burkavidins 1** (purple) (Fig. 7, Fig. 8). Two sequences, AMIN CIRC and RHOD SP01, did cluster with the clades. Their origin species, *Aminiphilus circumscriptus* and *Rhodonobacter sp. OR444*, were isolated from waste sludge and heavy metal polluted soil respectively.

The **dimeric avidins** clade contained all experimentally verified, dimeric avidins: rhizavidin (Rhizavd), hoefavidin (Hoefavd), and shwanavidin (Shwanavd). Also bradavidin II (Bradavd II) clustered into this clade, although its oligomerization is heterogenous. The origin bacterial species included nitrogen fixing, root nodular symbiotic, marine, marine and lake sediment, sponge surface flora, and plant pathogenic species (Fig. 8, Table Appendix A).

The **bradavidins 1** clade forms the only sister clade to the **dimeric avidins**. Rhodavidin (Rhodavd) and bradavidin I (Bradavd I) both contain the brad-tag sequence and were the defining sequences in this clade. Origin bacteria were again root nodular symbiotic species closely related to *Bradyrhizobium diazoefficiens* with the exception of a widely spread multitrophic species (*Rhodopseudomonas palustris*) (Fig. 8, Table Appendix A).

The **burkavidins 2** clade contained only one verified avidin, burkavidin 2 (Burkavd 2) from *Burkholderia pseudomallei*, a human pathogenic species. Also, the sequence XANT CAMP from *Xanthomonas campestris*, has been suggested as an avidin. In addition, the clade contained several other sequences from human and plant pathogens, predominantly from the *Xanthomonas* genus, but also an endosymbiont of a sea slug, marine sediment bacteria, and antibiotic soil bacteria (Fig. 8, Table Appendix A).

The **fungal and streptavidins** clade contained streptavidin, streptavidin v1 and streptavidin v2 (Streptavd, Streptavd v1, and Streptavd v2), and tamavidin 1 and tamavidin 2 (Tamavd 1 and Tamavd 2). Although, the fungal tamavidins were not clustered together with the rest of the eukaryotic avidins, a coinciding result was obtained previously with principal component analysis in his thesis work by Tiwari in 2015. Aside from the wood decaying fungus *Cornucopiae pleurotus*, the Tamavd 1 and 2 were isolated from, the clade mainly included sequences from forest soil bacteria; and marine sediment bacteria (Fig. 8, Table Appendix A).

The **animal avidin** clade included the rest of the verified eukaryotic avidins in the study set. This list covered the avidin (AVD) (*Gallus gallus*), strongavidin (Strongavd) from the sea urchin (*Strongylocentrotus purpuratum*), zebavidin (Zebavd) from the zebrafish (*Danio rerio*), and xenavidin (Xenavd) from the clawed frog (*Xenopus tropicalis*). Albeit, Strongavd has not been reported in literature yet.

The **burkavidins 1** was the last clade clustered in the superclade of the verified avidins, yet did not itself include any reported and verified avidins. However, several of the leaves were from the same species as the sequences of the **burkavidins 2** clade. The **burkavidins 1** group could be assumed homologous to the reported Burkavd 1 sequence. The clade contained sequences from the *Burkholderia* genus, as well as, other pathogenic bacteria, plant symbiotic bacteria, polluted lake sludge dwelling bacteria, marine species, sponge surface flora, and an endosymbiotic bacteria of parasitic nematodes (Fig. 8, Table Appendix A).

The direct sister superclade for the verified avidins consisted of three distinctly identifiable clades: **bradavidins 3**, **legavidins**, and **metavidins**. The **bradavidins 3** clade was defined by large group of symbiotic soil bacteria, including a plant endobiotic species; pathogenic bacteria; and finally an acidophilic decomposing forest soil bacterium (Fig. 7, Table Appendix A). The **legavidins** clade consisted of mainly human pathogenic *Legionella* species, but included also two endosymbiotic bacterial species of the sea slug and porous coral (Fig. 7, Table Appendix A). Meanwhile, the **metavidins** clade was small and included only three sequences, all from opportunistic aquatic pathogens of the *Methylomonas* genus (Fig. 8, Table Appendix A).

The last larger clade consisted of two smaller, clades with sequences from heterogenic bacterial species. The first clade, **β6 avidins**, was defined by highly unusual sequence in the β6-strand, and the second clade, **extended avidins** contained a C-terminal fusion in their sequences. The **β6 avidins** included aquatic bacterial species, a fish pathogen, and soil bacteria from oil contaminated sample (Fig. 8, Table Appendix A). Meanwhile, the **extended avidins** contained several pollutant and toxin degrading species, an acidophilic soil bacterium, an aquatic bacterium species isolated from a sepsis patient, a marsh grass and a seaweed microflora participating species, marine bacterioplankton species, and a probable bacteriosymbiont (Fig. 8, Table Appendix A).

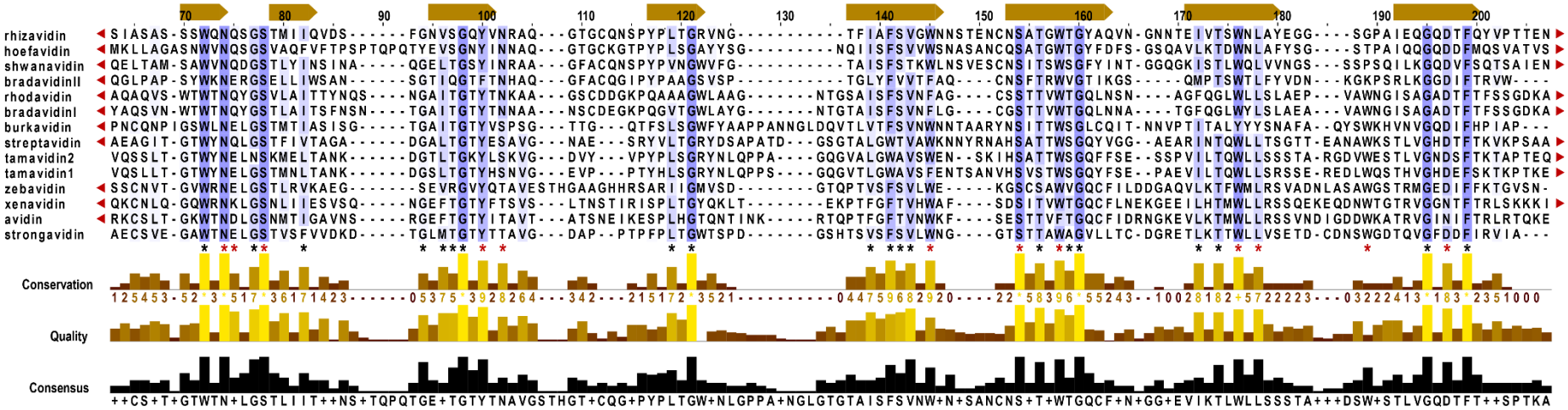


Figure 5. Structural multiple sequence alignment of the verified avidins set. Red triangles mark sequence continuation, brown arrows the β-strand positions in streptavidin, and asterisks the biotin-binding (red) or conserved positions (black). The alignment was produced with EXPRESSO, visualized with JalView software, and visually edited in InkScape.

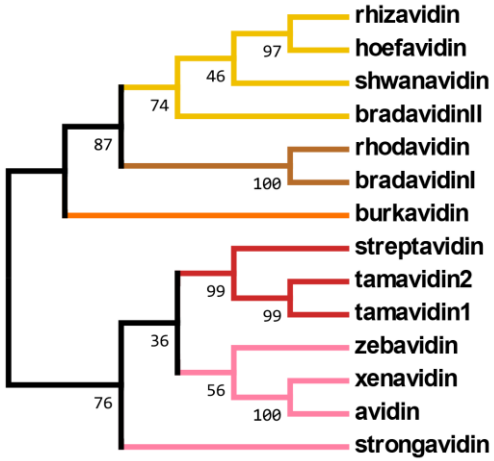


Figure 6. Phylogenetic cladogram tree of the verified avidin sequences. The clades are coloured to match the identified subgroups of the putative avidins cladogram (Fig. 5). The tree was computed and visualized in MEGA6.

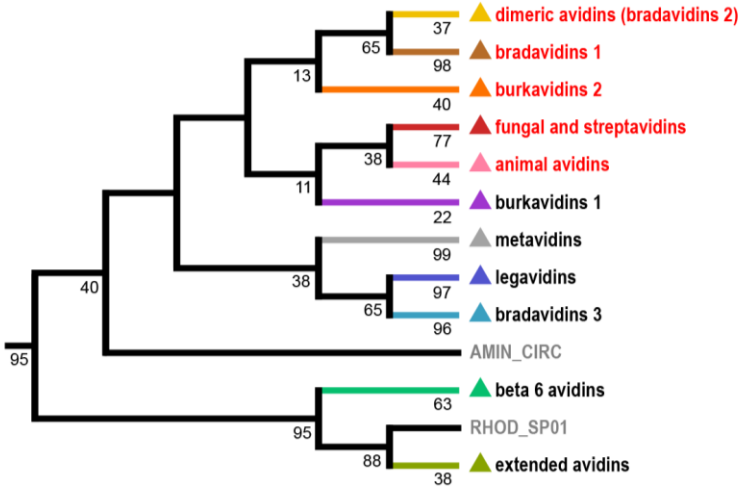


Figure 7. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with collapsed subgroups. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

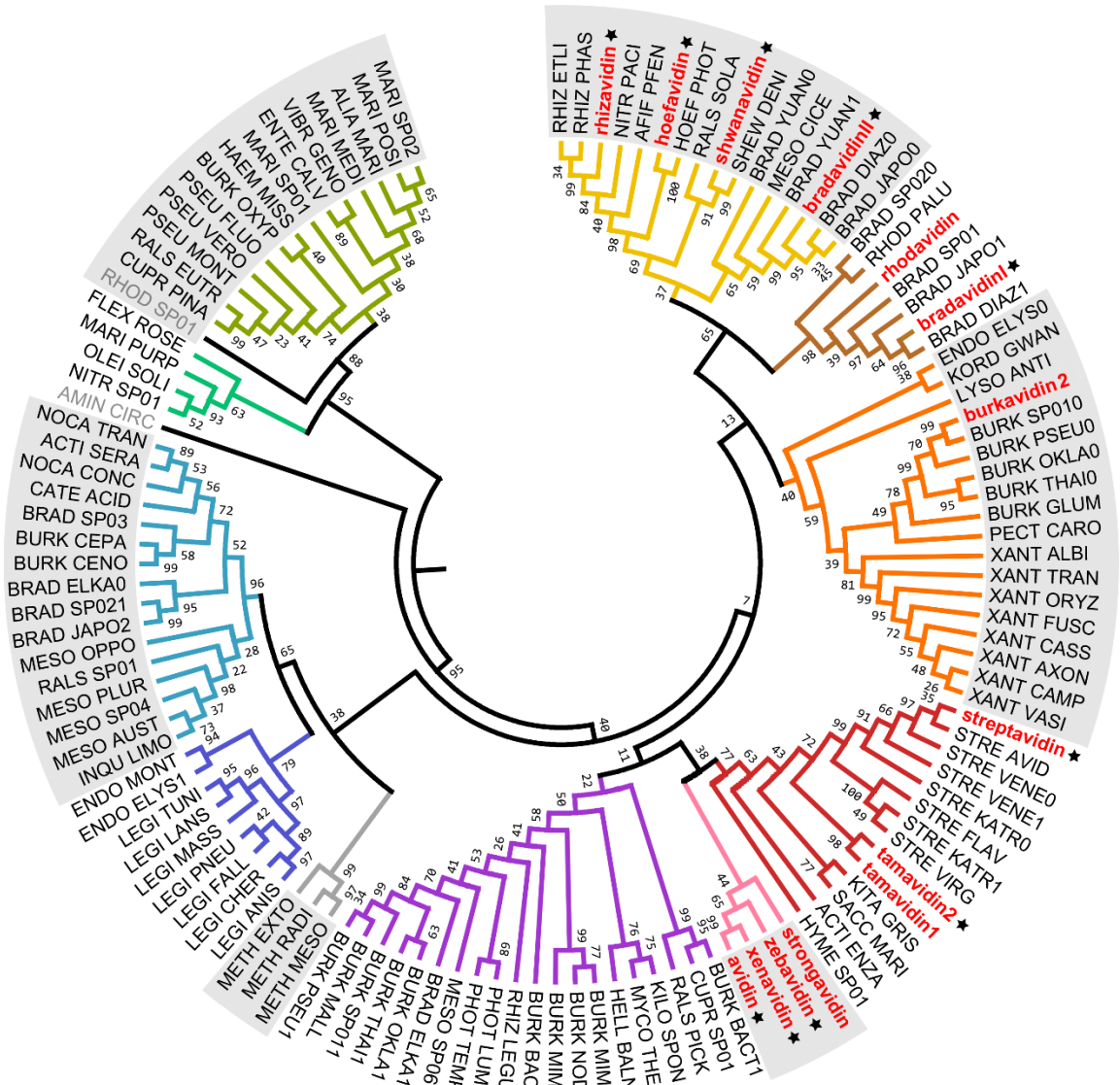


Figure 8. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences. The clades that were considered their own subgroups among avidin family are highlighted with different colours: gold for **dimeric avidins**, brown for **bradavidins 1**, orange for **burkavidins 2**, red for **fungal and streptavidins**, pink for **animal avidins**, purple for **burkavidins 1**, grey for **metavidins**, aquamarine for **legavidins**, powder blue for **bradavidins 3**, green for **β6 avidins**, and olive for **extended avidins**. Red node names mark the experimentally verified avidins, grey node names mark the sequences that did not fit any subgroup, and the star denotes the sequences with a PDB structure available. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

5.3.2 Sequence footprints in clades

The distinct clades were visualized as MSAs (Fig. 20–23) to further evaluate the differences in sequence between the groups. The clade alignments are taken from the full set MSA (data not shown due to size limitations). Furthermore, tables of the cladewise changes in 32 key positions, including 12 biotin-binding positions and 20 generally conserved positions, were collected (Table4, Table5). The AA positions in the cladewise MSAs will be numbered in the text by the corresponding Streptavd positions for clarity of comparison.

Dimeric avidins

The **dimeric avidins** clade (Fig. 9 and Fig. 20 for the phylogeny and MSA, respectively) shared 18 fully conserved AA positions, seven of which were homologous with the biotin-binding residues of Streptvd Asn23, Ser27, Ser45 (substituted to Asn in dimeric avidins), Trp92, Trp108, Leu110, and Asp128 (Table 4, Fig. 20). Further eight were also conserved between all verified avidins Trp21, Gly26, Gly41, Gly58, Thr90, Gly94, Gly126, Phe130 (Table 4, Fig. 20). The rest three of the conserved residues were an Ala at position 47 and Cys in both 49 and 86 positions.

While not fully conserved, rest of the biotin-binding residues had analogues in the clade's sequences, as well, with the exception of Trp120. While two sequences, BRAD YUAN1 and MESO CICE, had conservative Trp120Phe substitutions, most of the sequences contained a Pro at this position. This feature is well documented in dimeric avidins. Two sequences, AFIF PFEN and RALS SOLA, did not show any possibly compensating AA near the position.

Most of the other positions, generally conserved in avidins, were substituted conservatively, as well. Val31, Leu39, Leu56, Trp75, Leu78 and Ile104 positions were hydrophobic throughout the dimeric avidins sequences. However, positions Gly37, Thr40, Val77, and Ser93 contained few non-conservative substitutions: namely Gly37Trp in verified Hoefavd and HOEF PHOT; Gly37Asp in NITR PACI; Thr40Ala in BRAD YUAN0; Thr40Leu in MESO CICE; and Val77Thr in verified Shwanavd, SHEW DENI and RALS SOLA. Position, Thr76 on the other hand, showed more non-conservative substitutions, Thr76Ala was present in RALS SOLA; Thr76Tyr in BRAD YUAN0 and MESO CICE; Thr76Val in both Bradvd II and three sequences closest to it in the clade. Finally, Ser93Val substitution was present in the Bradvd II and the aforementioned three sequences closest to it.

There were also some notable, yet not completely conserved changes that were at the other positions. Positions 19 and 25 contained only polar AA residues, a feature unique for dimeric avidins. ProTyrPro sequence spanned positions 53 to 55 and interestingly this change

is also seen in Tamavd 2. Polar or negatively charged amino acids were present at the position 109 while this position was strictly hydrophobic in the rest of the verified avidins. While, the **dimeric avidins** showed predominantly aromatic residues at the position 112, most of the verified avidins contained a Ser, Arg or Asn residue at this position. Finally, at the position 125, the **dimeric avidins** featured a polar residue, while Val is common among the other verified avidins and is considered important for the tetramer integrity.

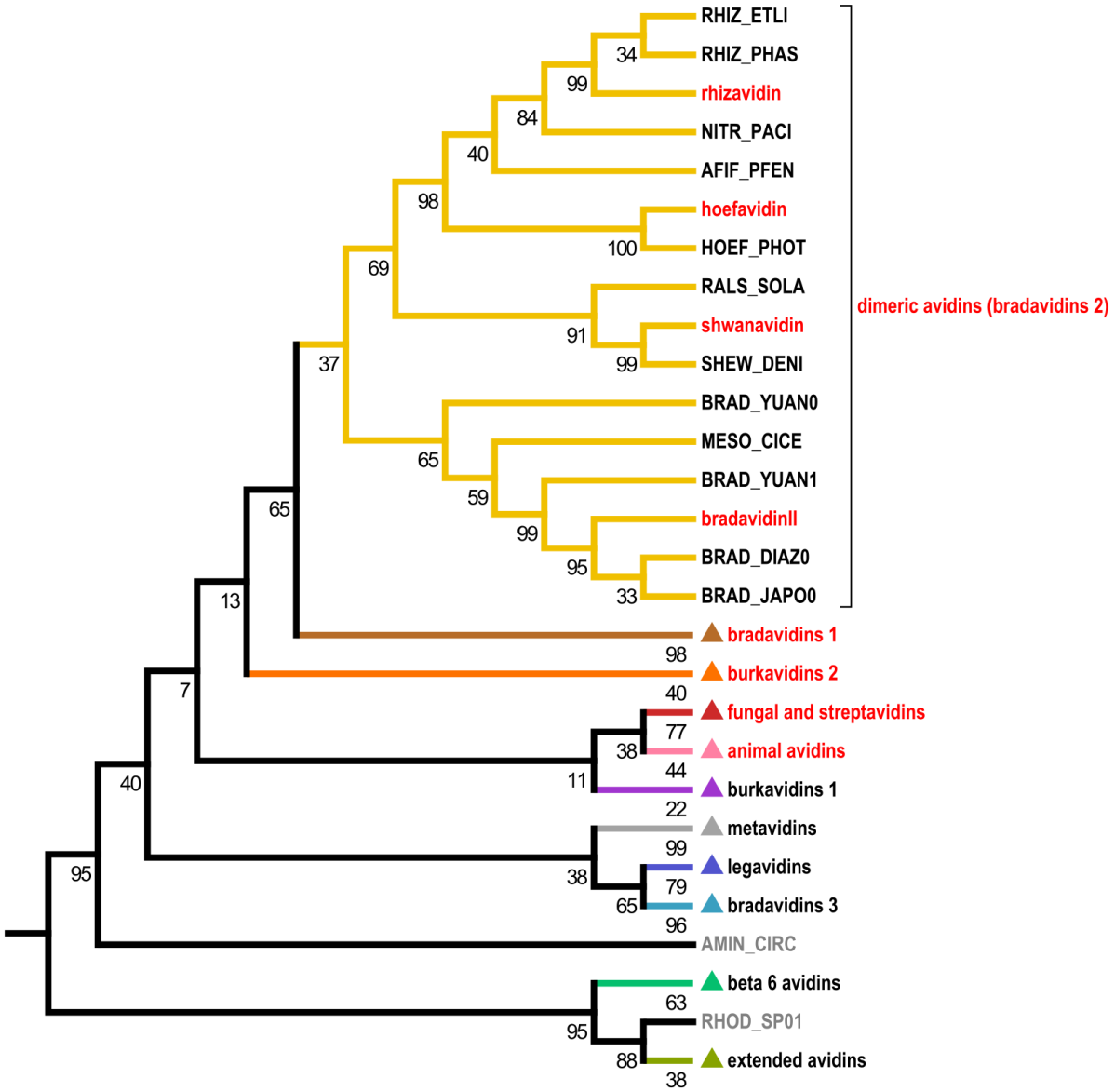


Figure 9. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the dimeric clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

Bradavidins 1

Bradavidins 1 group showed high conservation and was a rather small clade. It was separated from the **dimeric avidins** for a chance to explore the differences between the oligomeric states (Fig. 10 and Fig. 20 for the phylogeny and MSA, respectively). Most biotin-binding positions were perfectly conserved between Streptavd and the **bradavidins 1** clade sequences. Two positions contained conservative substitutions: Ser45 was substituted with Asn, a change also present in the **dimeric avidins**; and Trp79 was changed to Phe, substitution that is unique for **bradavidins 1** and tamavidin 1 sequences (Table 4, Fig. 20). Half of the 20 positions conserved generally in avidins were also conserved among the **bradavidins 1** sequences. Two positions contained conservative substitutions, when compared to Streptavd: Trp75 was substituted with Phe, similar to AVD; while Ile104 substitution to Phe was not observed in verified avidins outside this branch (Table 4, Fig. 20).

The rest of the generally conserved positions contained some variation. However, the conserved hydrophobic positions Val31, Leu39, Leu56 and Leu73 contained only hydrophobic residues in this branch, as well. In addition, Thr76 was either conserved or the position contained conservative substitution to Ser. Only the Thr40 and Thr106 positions contained

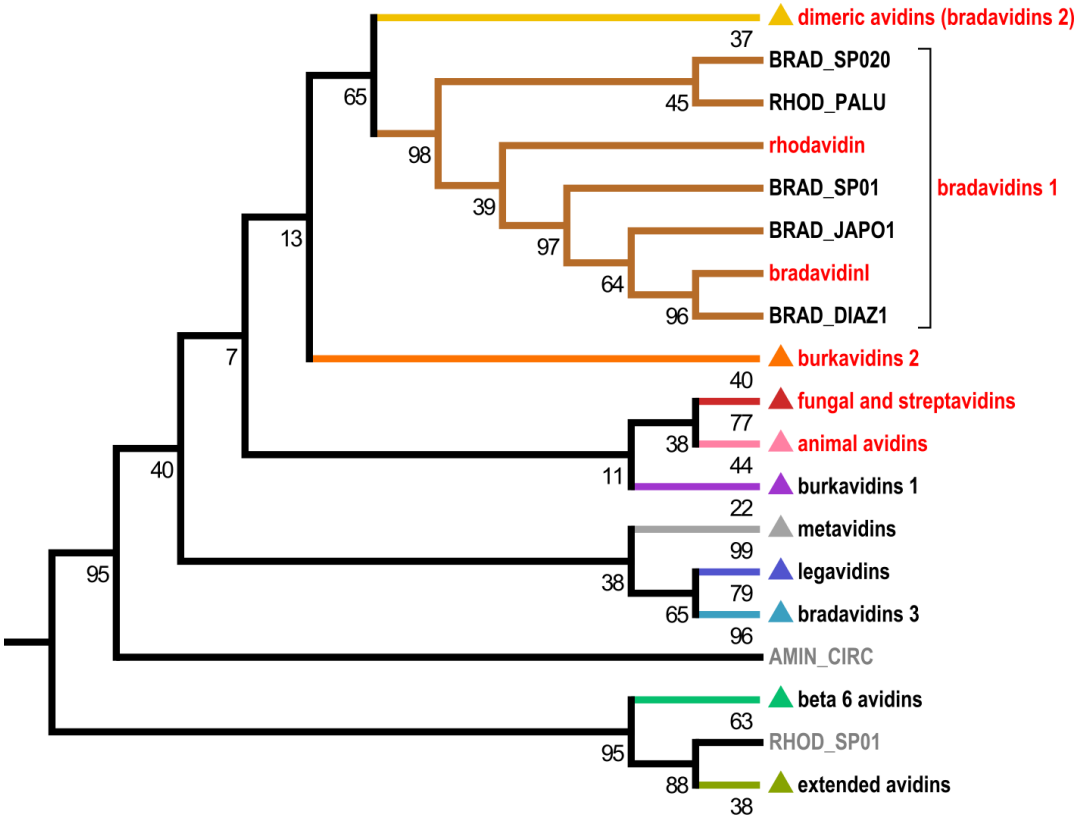


Figure 10. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the bradavidins 1 clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

some non-conservative variation. Thr40 position had two non-conservative substitutions to Ala, in RHOD PALU and BRAD JAPO1 sequences. Meanwhile, the Thr106 position was not conserved at all with Thr106Gly as the dominating substitution with one substitution of Thr106Val in BRAD SP020 sequence.

In addition to these, some uncommon conservation was present in few position. Position 19 contained a Trp residue, a change so far unique for the verified Rhodavd and Bradavd I sequences; position 25 featured an aromatic Tyr residue, while generally this position contains a Leu or a polar or a charged residue in the **dimeric avidins**; Ala occurred at position 47, similar to the **dimeric avidins**; two Cys capable of forming a C–C bridge were present at positions 48 and 87, again like in **dimeric avidins**; Val125, essential for tetramer integrity was substituted with Ala; and finally the sequences contained identical C-terminus spanning from Phe130 onwards.

Burkavidins 2

The MSA of **burkavidins 2** clade contained 12 strictly conserved amino acid positions (Fig. 11 and Fig. 20 for the phylogeny and MSA, respectively). Only one of these, Trp92, was among the common biotin-binding amino acids (Table 4, Fig. 20). Instead, 7 were homologous with the generally conserved positions, Trp21; Gly26; Gly37; Gly58; Trp75Phe; Val77; and Gly94 (Table 4, Fig. 20). Two Cys residues, likely participating in a C–C bridge, were strictly conserved at positions 14 and 96, similar to the verified avidins of animal origin. Position 19 was occupied by a Gly. Interestingly, aside from the sequences in the **dimeric** and **bradavidins 1** clades, all the verified avidins contain this Gly19. The final strictly conserved position, a His residue at the position 122, was completely unique for the Burkavd 2 among the verified avidins.

Most of the other biotin-binding positions showed conservation as well. For the positions Tyr43, Ser45 and Trp79 only conservative substitutions were present. While not fully conserved in the whole branch, the rest of the biotin-binding positions remained conserved aside from two outgroup sequences: ENDO ELYS0 and KORD GWAN. At the positions Asn23, Gln24, and Ser27, these two sequences contained changes Asn23Thr, Gln24Ser or Gln24Asn respectively, and Ser27Ala. Rest of the sequences showed conservation at these positions. Furthermore, positions Ser88 and Trp108 showed non-conservative substitutions of Ser88Ala and Trp108Asp in the KORD GWAN sequence. Also, otherwise strictly conserved positions in this clade, Trp120 and Asp128, were replaced by Phe and Thr, respectively, at positions 119 and 128 in ENDO ELYS0. Interestingly, position Leu110 has been strictly conserved aside

from the Burkavd 2 featuring aromatic Tyr at this position. Even within this clade, sequences from *Burkholderia* species contained aromatic Tyr or Phe residue at position 110, while rest of the sequences featured Leu, with the exception of ENDO ELYS0 containing Leu110Trp substitution.

The generally conserved hydrophobic locations, Val31, Leu39, Leu56, Leu73 and Ile104, remained conservatively hydrophobic throughout the sequences. Positions Thr40 and Gly41 contained homologues except for substitution to Val in both locations in the ENDO

Figure 11. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the burkavidins 2 clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

ELYS0 sequence. Thr76 position was conserved or featured conservative substitution to Ser except for the KORD GWAN sequence, which showed a substitution of Thr76Val. Similarly, Thr90 position contained one non-conservative substitution to Ala, but in the sequence, LYSO ANTI. The Ser93Val substitution was present in the sequence XANT TRAN and Thr106Ala substitution was seen in all the *Burkholderia* sequences within the branch, as well as, in the PECT CARO sequence. Lastly, there was a substitution of Gly128Asn in half of the sequences, namely the sequences from the *Xanthomonas* species and the KORD GWAN sequence.

As with the previously mentioned clades, **burkavidins 2**, contained several positions that showed more conservation than in general within avidins. Position 16 contained a polar Asn or Ser followed mainly by a Pro residue. In the β4-strand, there were two positions, a Pro at the position 54, and an aromatic residue at the position 55, conserved unlike in the verified avidins in general. Similarly, the position 59 contained a Trp almost without change. Position 102 at the beginning of the β7-strand contained mostly a Pro residue, a change that is seen in the verified Tamavd 2, as well. There was also an additional aromatic residue at the position 118 near the biotin-binding Trp120 and finally at the C-terminus, there was a Pro residue at the position 132 after the Phe130.

Fungal and streptavidins

In the **fungal and streptavidins** clade, the MSA showed 26 fully conserved positions (Fig. 12 and Fig. 21 for the phylogeny and MSA, respectively). Eight of these were biotin-binding positions: Ser27, Tyr43, Ser88, Trp92, Trp108, Leu110, Trp120 and Asp128 (Table 4, Fig. 21). Furthermore, 8 of the fully conserved positions within the **fungal and streptavidins** branch were also conserved generally among avidins: Trp21, Gly37, Gly41, Gly58, Ser93, Gly94, Gly126 and Phe130 (Table 4, Fig. 21). The last 10 of the fully conserved positions included several additional glycines at positions 19, 48, 70 and 74; two Tyr residues at the positions 22 and 54; an Asn residue at the position 81; a Gly at 95; a Leu at 109; and finally a Thr at 123.

The rest of the biotin-binding residues, however not fully conserved, still contained homologues with only conserved substitutions present. The only conservative substitution that has not been found in an experimentally verified avidin sequence was Trp79Tyr. This substitution was present in three sequences: STRE FLAV, STRE KATR1 and STRE VIRG.

Out of the remaining generally conserved positions, the hydrophobic locations Val31, Leu39, Leu56, Leu73, Val77 and Ile104 remained hydrophobic in **fungal and streptavidins**, as well. Trp75 was conserved aside from the sequence HYME SP01, which contained substitution to Phe. This position features a Phe residue in all other the verified avidins. The

position of Gly26 contained two non-conservative substitutions: Gly26Asn in the confirmed Tamavd 2 sequence, and Gly26Gln in the sequence ACTI ENZA. Four generally conserved Thr residues, Thr40, Thr76, Thr90 and Thr106, were mostly conserved, but featured a few hydrophobic substitutions each. Thr40Val substitution was present in the STRE KATR1, STRE VIRG and ACTI ENZA sequences. Thr76 was changed to Ala in the two fungal verified avidins, Tamavd 1 and Tamavd 2, and to Val in the HYME SP01 sequence. A change of Thr90Ala was seen in three *Streptomyces* sequences: STRE FLAV, STRE KATR1 and STRE VIRG. While, the substitution Thr106Ala occurred in the KITA GRIS sequence.

A number of positions showed also conservation within this clade, aside from the generally conserved positions. Position 17 featured Ile residue except for the Tamavd 1 and

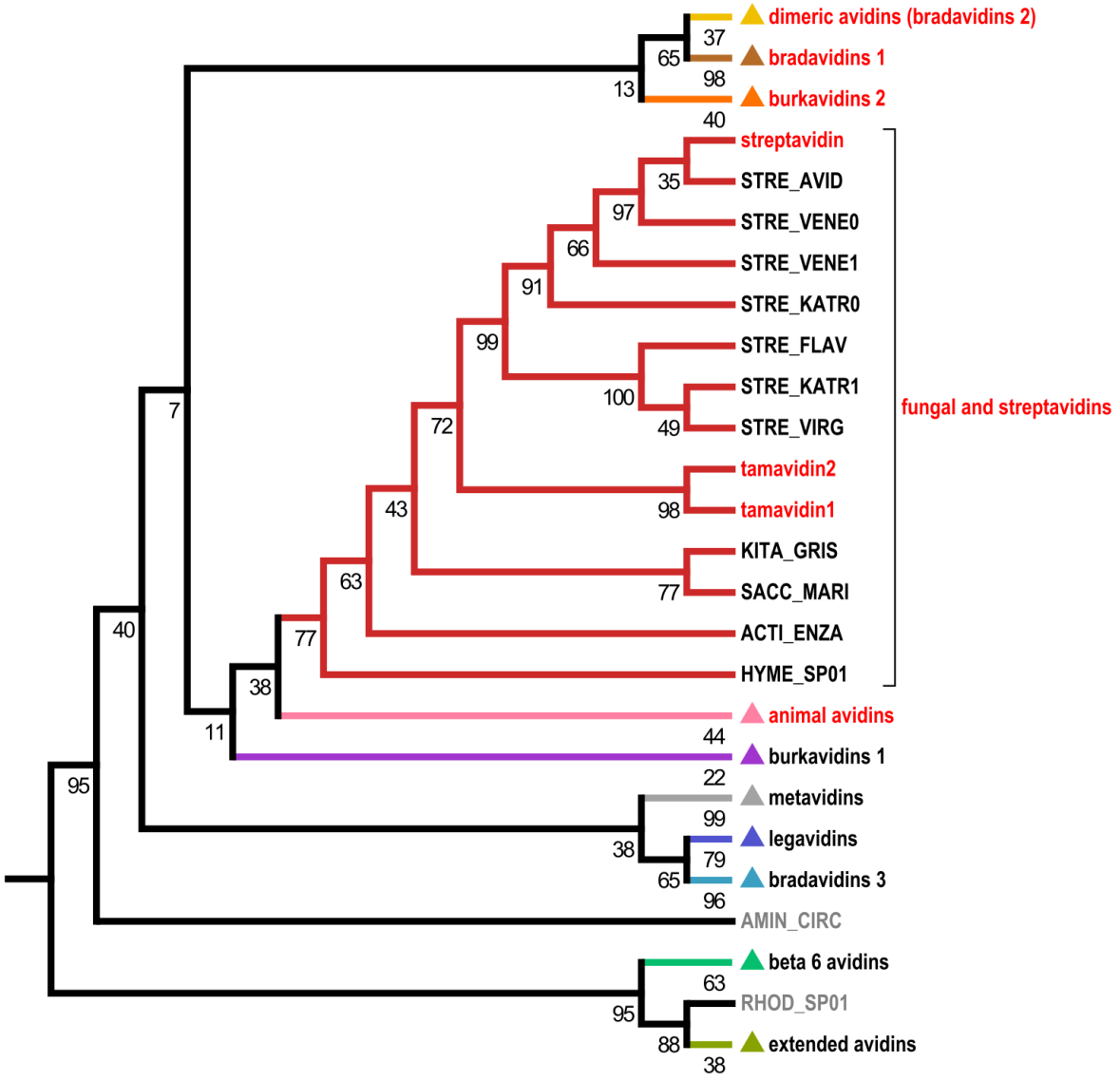


Figure 12. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the fungal and streptavidins clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

Tamavd 2. Positions 18 and 20 contained a Thr residue in the *Steptomyces* species and the two fungal avidins. However, the Thr residue at these two positions is rather uncommon in the other confirmed avidins. Position 25 featured almost uniformly a Leu residue and interestingly this feature is also common in all but the **dimeric avidins** and **bradavidins 1** in the **verified avidins** superclade. Right before the perfectly conserved Gly48, at the position 47, there was Val with the exception of sequence KITA GRIS. This position is hydrophobic in the verified avidins aside from the Burkavd 2. At the end of β4-strand, there was a highly conserved Arg residue at the position 59. This curious change was followed by a Tyr residue. A similar conserved Tyr was present also at the position 96. This location was hydrophobic or contained a Cys residue in the verified avidins. Positions 118, 121, 122, and 132 contained either polar or charged residues. Verified animal avidins, showed polar and charged residues at the positions 118, 122 and 132, as well. Lastly there was a Pro residue at the position 132. Unlike the other clades that contained verified avidin sequences, fungal and streptavidins did not feature any conserved Cys residues.

Animal avidins

The **animal avidins** clade, was small and showed strict conservation at 27 positions (Fig. 13 and Fig. 21 for the phylogeny and MSA, respectively). Nine of these positions were biotin-binding: Asn24, Ser27, Tyr43, Ser45Thr, Trp79, Ser88, Trp108, Leu110, and Trp120 (Table 4, Fig. 21). Eleven more were generally conserved among the verified avidins: Trp21, Gly26, Gly41, Gly58, Trp75Phe, Val77, Gly94, Ile104Leu, Thr106, Gy126 and Phe130 (Table 4, Fig. 21). Two of the remaining conserved positions were Gly residues, namely Gly19 and Gly100. Gly19 is conserved among the verified avidins aside from the **dimeric** and **bradavidins 1** clades. Also the Val47, Ser112, and Thr123 were conserved and the positions contained similarity in the **fungal and streptavidins** as well. Lastly a Cys at the position 15 was strictly conserved. This residue has been shown to form a C–C bridge with another Cys that shows slight variation in location from position 96 to 99.

The rest three of the biotin-binding positions were conservatively conserved, as well. Trp92 was substituted to a Phe in the AVD sequence and the position Asp128 contained an Asn residue in Xenavd and AVD. Finally, the Gln24 was substituted to a Lys residue in Xenavd and an Asp in AVD sequences.

In addition, the generally conserved positions showed also high conservation. As with the previous clades, the positions Val31, Leu39, Leu56 and Leu73 remained uniformly

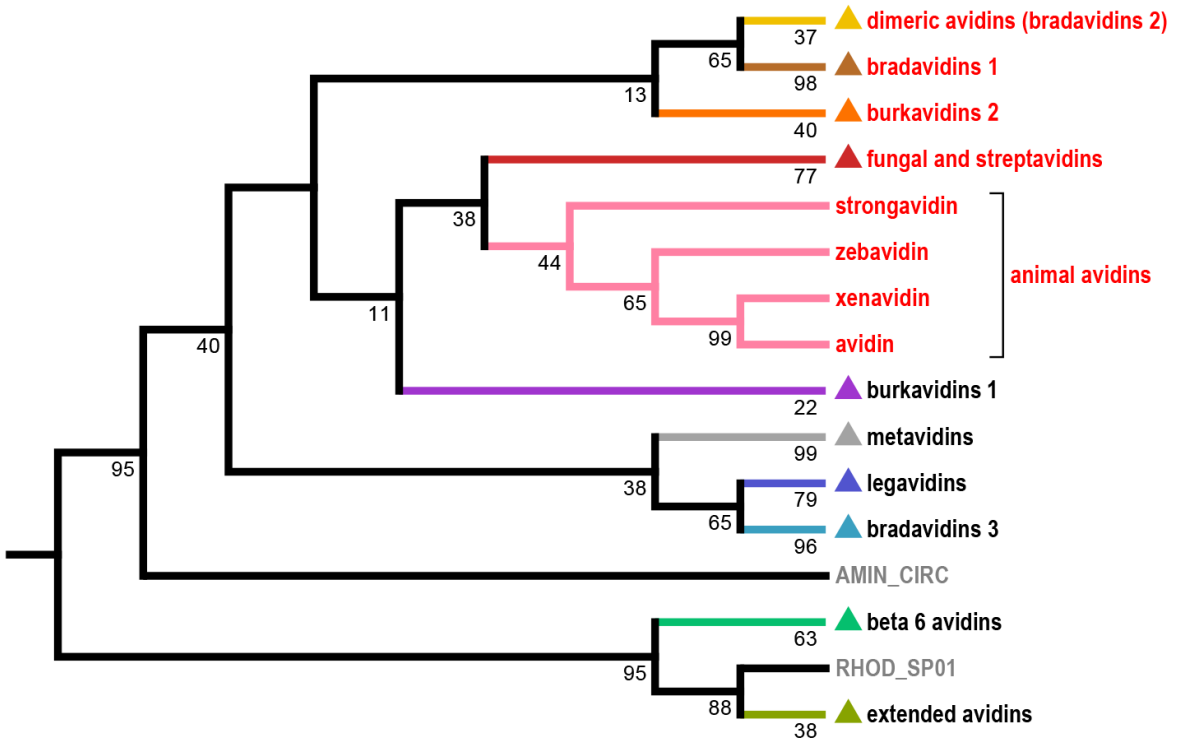


Figure 13. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the animal avidins clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

hydrophobic. Gly34 while not perfectly conserved was either present or shifted in the Zebavd. The Thr40 was substituted to an Arg residue in Zebavd, while the Thr76 and Thr90 were either conserved or contained a conservative substitution to Ser. However, the Ser93 was changed to a hydrophobic Val or Ala residue in Zebavd and Strongavd sequences respectively.

Some other positions showed further similarities. Location 72 showed a polar Ser or Thr residue, while most other avidins featured a hydrophobic residue at this location. The position 91 was hydrophobic, similarly to the **bradavidins 1** clade sequences. A Glu or Gln residue was present at the location 102, similar to the verified Tamavd 1 sequence. The position 124 contained an Arg residue except for the Strongavd, and finally another positively charged residue was present at the location 132. The position 132 features positively charged residues in the **fungal and streptavidins** clade, as well.

Burkavidins 1

In the second clade dominated by the species of the *Burkholderia* genus, **burkavidins 1**, there were only 9 perfectly conserved positions (Fig. 14 and Fig. 21 for the phylogeny and MSA, respectively). However, when 3 outgroup sequences, BURK BACT1, CUPR SP01 and RALS PICK, were excluded, the amount rises to 14. Two of the perfectly conserved residues were the biotin-binding positions, Trp108 and Trp120 (Table 4, Fig. 21), and the 5 positions conserved

aside from the outgroup sequences contained three more biotin-binding positions, Asn23, Ser27 and Trp92. Six of the strictly conserved positions were among the generally conserved residues: three Gly residues at positions 41, 58 and 94; aromatic Trp21 and Phe75; and lastly the hydrophobic Val77 residue. When the positions conserved aside from the three outgroup sequences were considered, two more generally conserved positions, Gly126 and Phe130 (Table 4, Fig.

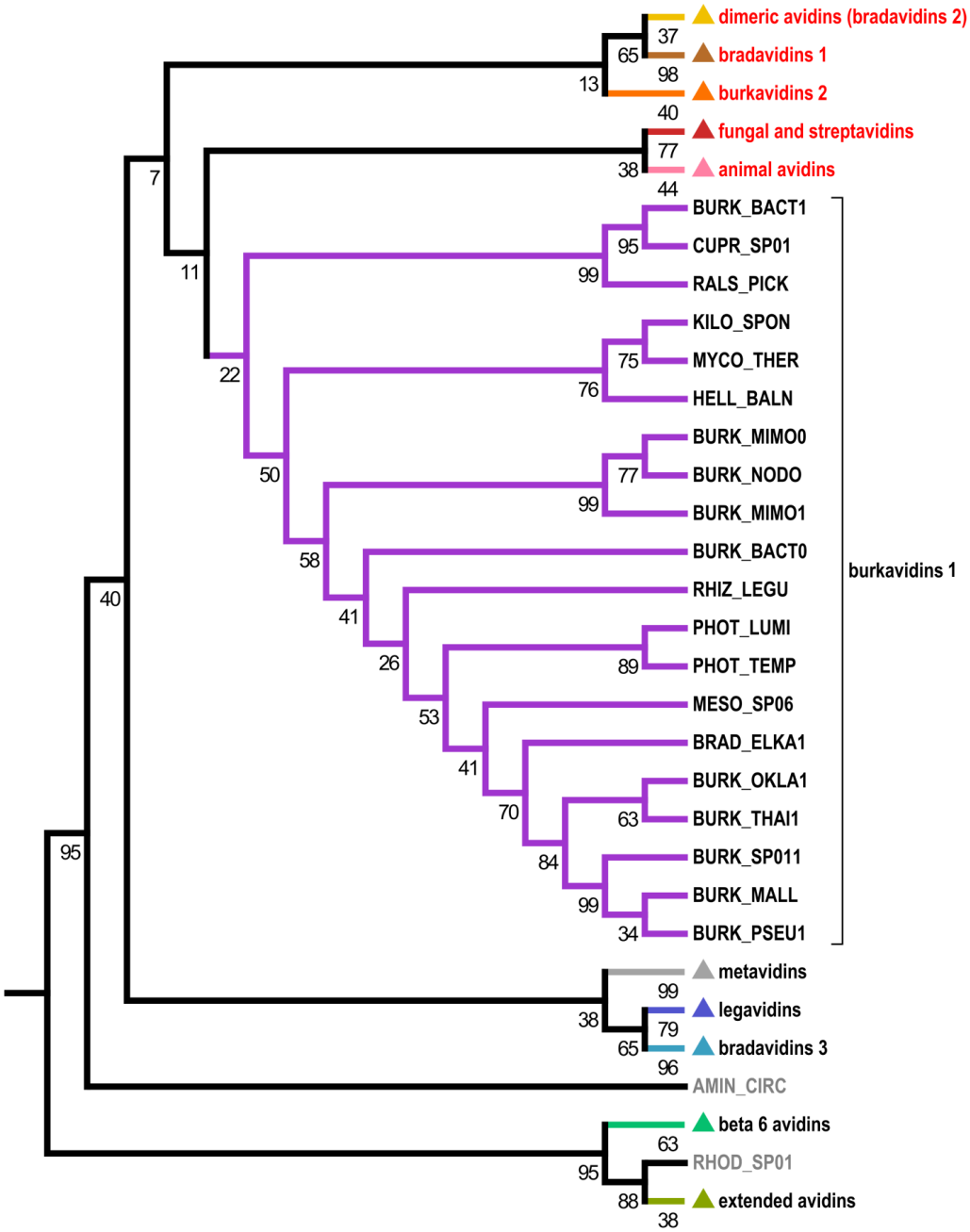


Figure 14. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the burkavidins 1 clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

21), showed strict conservation. In the outgroup sequences, changes Gly126Asn and Phe130Trp were observed.

Furthermore, the other biotin-binding positions were mainly conserved, with only two exceptions: Ser88 and Leu110. Ser88 was non-conservatively substituted to an Ala and Leu110 conservatively to a Met residue in 11 of the clade's 21 sequences. Other changes were present in some individual sequences. The outgroup sequences showed some unusual changes: Glu24Arg, and Asp128Ala. Glu24 was also substituted to a Gly residue in the sequence KILO SPON.

Even the generally conserved positions were mostly preserved. Only notable change at the hydrophobic positions was the change of Leu56Thr in the sequence RALS PICK. Two glycines, Gly26 and Gly37, contained some non-conservative substitutions and the Thr76 and Thr93 positions showed hydrophobic substitutions. These changes were present in the outgroup sequences and another set of three sequences: KILO SPON, MYCO THER, and HELL BALN.

Only few locations outside of the biotin-binding or generally conserved positions showed conservation in the **burkavidins 1** clade. The position 17 was uniformly hydrophobic aside from the three outgroup sequences. Position 19 contained a Gly residue, which were common in the verified avidins aside from the **dimeric** and the **bradavidins 1** clades. Position 123 was strictly hydrophobic, similar to the two verified bradavidins, Bradavd I and Bradavd II. The Val125 was conservatively substituted to an Ala residue again with the exception of the outgroup sequences, which contained a substitution of Val125Ser.

Metavidins

The **metavidins** clade contained only three sequences and thus contained a high amount of fully conserved positions (Fig. 15 and Fig. 22 for the phylogeny and MSA, respectively). However, only one of these, Trp92, was within the biotin-binding positions (Table 4, Fig. 22). Instead, a total of 14 were among the generally conserved positions: Gly37, Thr40Ser, Gly41, Leu56, Gly58, Leu73Ile, Trp75Phe, Thr76Val, Val77Thr, Thr90, Ser93Thr, Gly94, Ile104 and Phe130 (Table 4, Fig. 22). Most of the other strictly conserved positions were spread among the spans corresponding to the β-strands in verified avidins.

The biotin-binding positions were not well conserved either. Two of the three sequences, METH MESO and METH RADI, were truncated from beginning missing the two first β-strands. Hence the possible homologues for Asn23, Glu24 and Ser27 were not present in these sequences. In the one sequence without truncation, Asn23 and Glu24 were present, but Ser27 was substituted to Ala. There was a substitution Tyr43Cys in the non-truncated METH

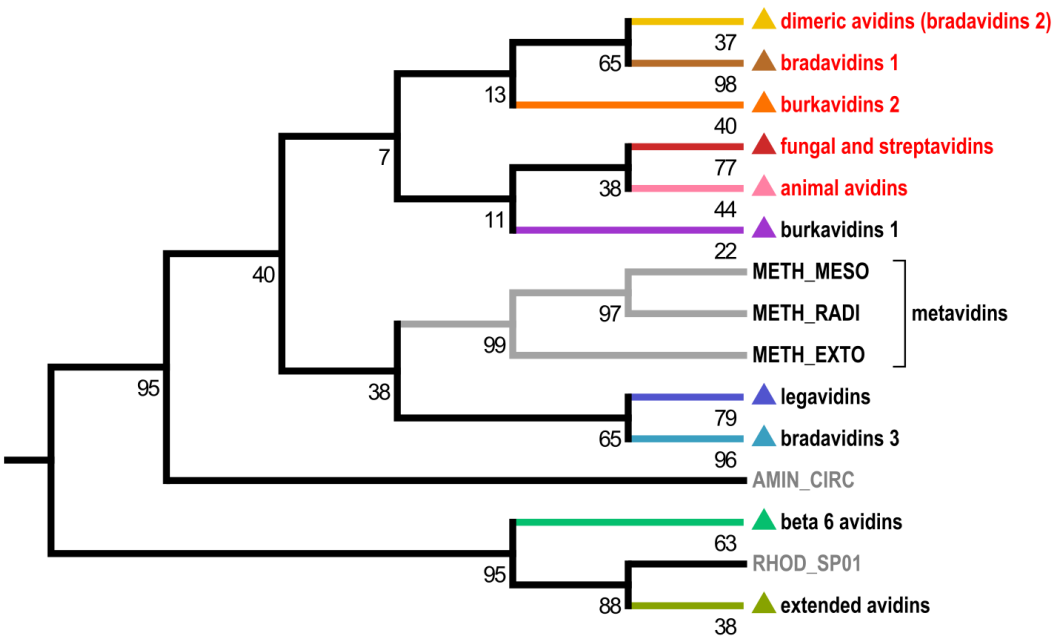


Figure 15. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the metavidins clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

EXTO sequence. Ser45 was replaced by Ala or Pro, but there was a polar Thr residue at the preceding position 44. For Thr79 there was no homologue present and Ser88 was also substituted, yet conservatively, to Glu or Thr. While Leu110 position contained hydrophobic residue, there was no homologue present for the Trp120. Lastly, the Asp128 position contained conservative substitution to Thr or non-conservative substitution to Ala in the sequence METH EXTO.

For the generally conserved positions, again the three positions in the span of two first β-strands were conserved in the sequence METH EXTO with the segment intact. Thr106 was substituted to a hydrophobic Ala or Leu residue. Finally, there was a change of Gly126Glu in the two truncated sequences: METH MESO and METH RADI.

Legavidins

In the **legavidins** clade, there was 13 fully conserved positions in the MSA (Fig. 16 and Fig. 22 for the phylogeny and MSA, respectively). Four positions of these were the biotin-binding residues Asn23, Ser27, Tyr43Phe and Ser45Thr (Table 4, Fig. 22) and five were among the generally conserved ones, Gly26, Val31Leu, Gly58, Thr76Ser and Gly94 (Table 4, Fig. 22). Last four were neither, but formed a six AA long strictly conserved sequence PheThrThrAlaValAla at the position homologous to Streptavd positions from 43 to 48.

Rest of the biotin-binding positions were mostly preserved. However, Glu24 position contained a range of polar residues including Glu, Gln, Thr or Lys. The substitution of Trp79Tyr

followed by a Pro residue was prevalent. In the ENDO MONT sequence, there was a possibly compensatory Phe residue at the position 77. The Ser88 was conserved, but in the sequences from the *Endozoicomonas* species, the Ser residue was moved by two AAs. Trp92 was not conserved, but the position features exclusively hydrophobic residues. Trp108 was conserved in the three sequences LEGI MASS, LEGI LANS and LEGI TUNI, but was substituted to a hydrophobic Ala in the sequence ENDO ELYS1 and to a polar Ser or Thr in the other sequences. The Trp120 was conserved in the two *Endozoicomonas* species, substituted to a Pro residue in the LEGI ANIS, LEGI CHER, LEGI FALL and LEGI PNEU sequences, and finally substituted to a hydrophobic Leu residue in the LEGI MASS, LEGI LANS and LEGI TUNI sequences. The Trp120Pro substitution was present in the verified **dimeric avidins**. Lastly, the position Asp128 was conserved or conservatively substituted to Asn.

The generally conserved locations were mainly present, but contained more changes in individual sequences than the clades with verified avidins. Especially two outgroup sequences, ENDO ELYS and ENDO MONT, showed some unusual changes. The Trp21 position featured either aromatic Phe or Tyr residue. The position Gly37 was either conserved of contained a

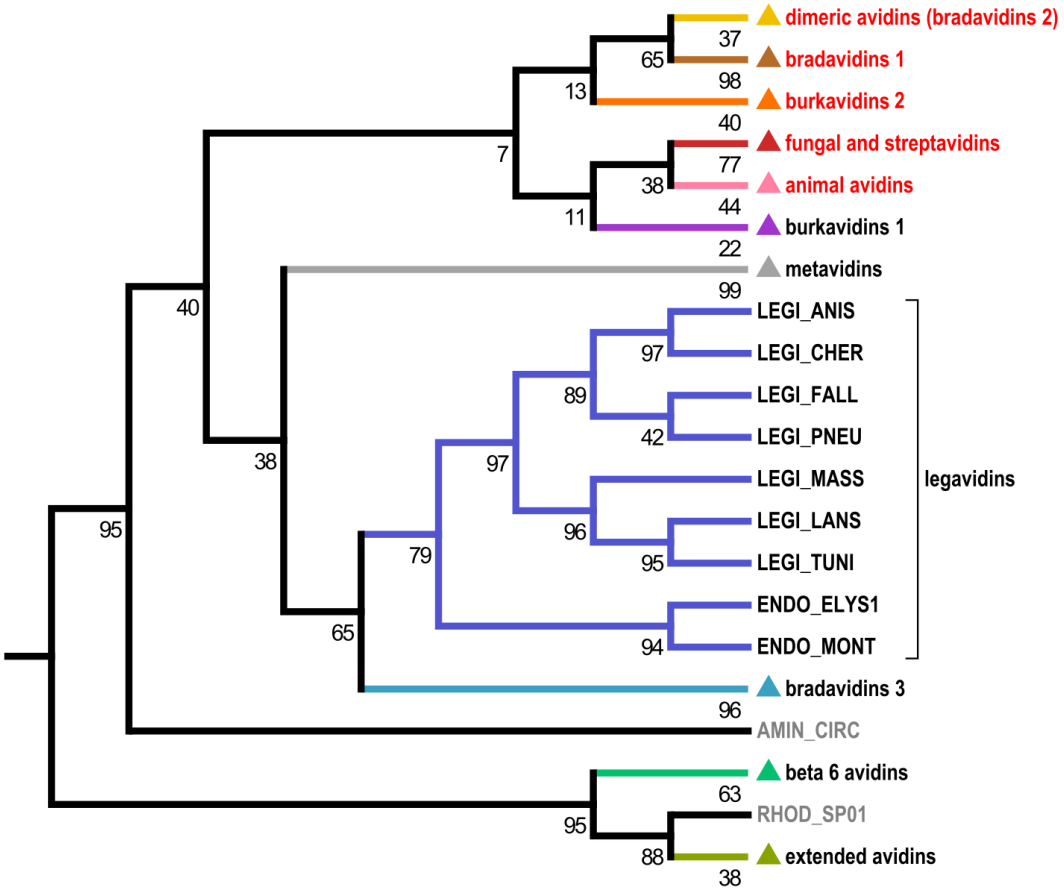


Figure 16. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the legavidins clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

polar substitution. There was a change of Leu56Cys in the ENDO ELYS1 sequence. Thr40 position contained an array of polar residues. The position of Trp75 was either conserved or contained a hydrophobic residue except in the ENDO MONT sequence, which featured an Asn residue. Thr90 was substituted with a hydrophobic residue. Meanwhile, the Ser93 was conserved with the exception of three sequences, LEGI MASS, LEGI LANS and LEGI TUNI. However, these sequences contained a perhaps compensatory Ser at the position 91. The change Gly126Thr was present in the two *Endozoicomonas* sequences, and lastly Phe130 was mostly substituted to a Tyr residue, similar to half of the **extended avidins** sequences. This change has not been reported in any of the verified avidins.

Some other conserved changes were present, as well. At the position 22, there was a Lys residue. An Arg, Lys or Asn residue was present at the position 25, similarly to the **dimeric avidins**. At the β2-strand, there was a strictly preserved Leu at the position 29. An additional Tyr preceding the Tyr43 was observed in most of the sequences. A Cys residue was present at the loop 3–4 and at the position 86 and a Pro residue was conserved at the position 55. These features are seen in the **dimeric avidins**, as well.

Bradavidins 3

The **bradavidins 3** clade contained 21 strictly conserved positions (Fig. 17 and Fig. 22 for the phylogeny and MSA, respectively), six of these being biotin-binding residues: Asn23, Ser27, Phe43, Thr45, Trp108 and Asp128 (Table 4, Fig. 22), and six more were among the generally conserved positions Trp21, Gly26, Gly41, Gly58, Gly94 and Phe130 (Table 4, Fig. 22). Rest of the strictly conserved positions were neither, but contained three often conserved residues: Trp17, Gly19 and Arg132. The loop 3–4 contained five perfectly conserved amino acids: Leu47, Asp49, Ser50, Phe52 and Gly54. Lastly there was a conserved Val residue at the position 111.

Most other biotin-binding positions were well conserved, as well. The Gln24, Trp92 and Trp120 residues contained homologues. The Leu110 was substituted with a Val residue or in sequences INQU LIMO, MESO AUST, and MESO SP04 with a Thr residue. For the residues Trp79 and Ser88 there was no clear homologue present.

Many generally conserved positions were preserved, as well, however there were some without a proper homologue, as well. The Trp75 contained an aromatic Phe. The Gly37 was substituted to a polar residue in almost half of the sequences and the Gly124 was uniformly substituted to either Ser or Asn. The Thr76 residue didn't show conservation.

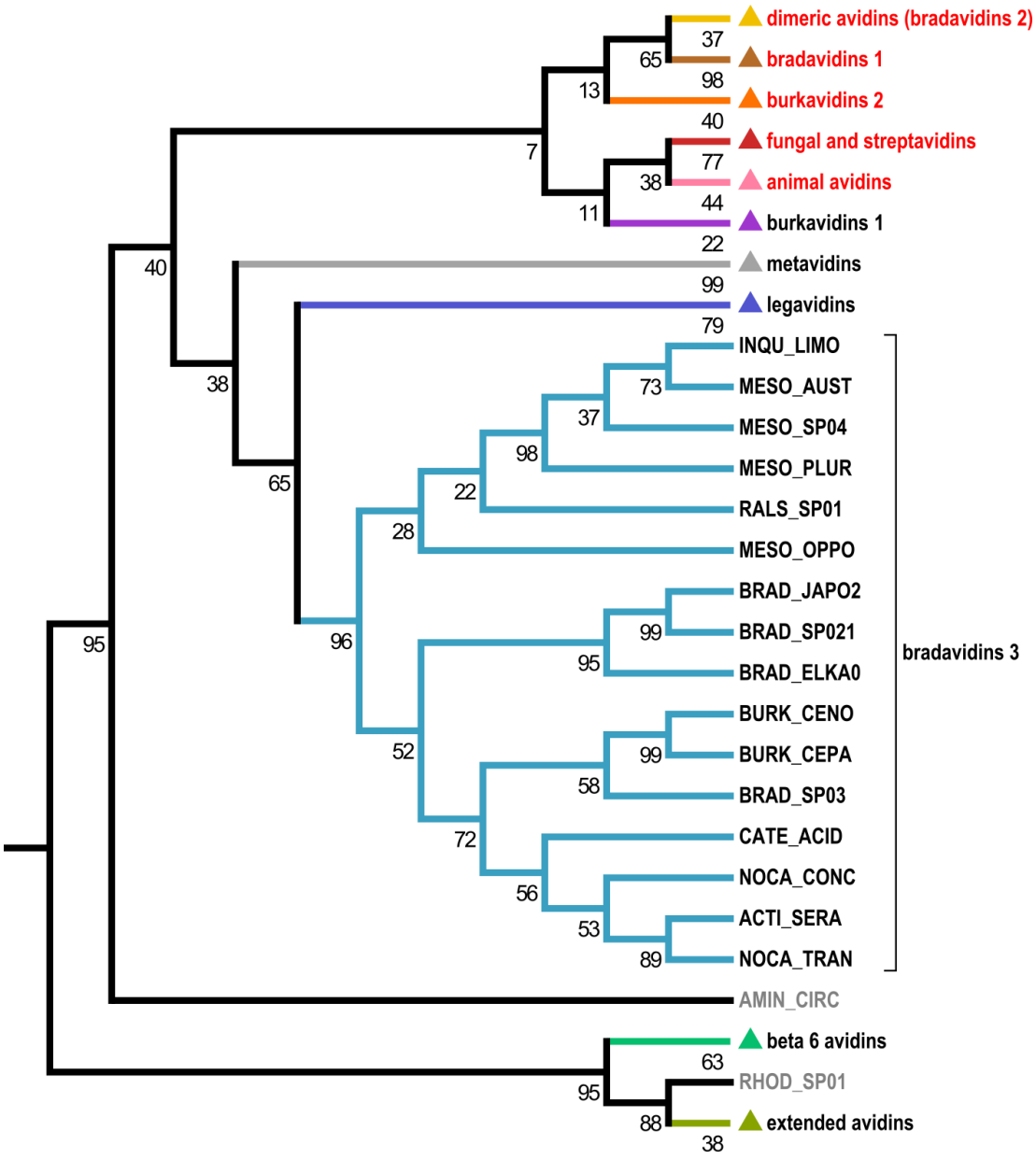


Figure 17. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the bradavidins 3 clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

As for other preserved locations, the position 29 was uniformly hydrophobic and the position 34 contained a polar Asp or Glu residue. There was a Gly residue at the position 62, and finally a conserved Cys residue in varying location in the loop 4–5. Another Cys residue was present in some sequences in the loop 5–6 or the β6-strand.

β6 avidins

The **β6 avidins** was a rather small clade and there were total of 36 fully conserved residues (Fig. 18 and Fig. 23 for the phylogeny and MSA, respectively). Eight of these residues were biotin-binding positions: Tyr43, Ser45, Trp79, Trp92, conservative substitution of Trp108Tyr,

Leu110, Trp120, and another conservative substitution of Asn128Ser (Table 4, Fig. 23). Nine more were among the generally conserved positions: Val31Ile, Gly37, Gly41, substitution of Leu56Val, Gly58, Trp75Phe, Val77, and Ser93 was shifted one AA further, Thr106, and Phe130 (Table 4, Fig. 23). The rest 19 of these strictly conserved residues were included the loop 3–4, which was almost fully conserved in the clade sequences. A Tyr was present at the position 54, similar to the **dimeric avidins**, Tamavd 1 and Tamavd 2. There were three conserved residues, Asp, Pro and Ser, in the loop 4–5. The rest of the conserved positions included: an Arg residue at the position 80, a Val between the Trp92 and Ser93 homologues, a Phe at 95, an Asp at 99, a Gly at 100, a Gln at 111, a Thr at 123, and finally another Thr at 129.

Three of the other biotin-binding positions were missing in the sequence FLEX ROSE, as it contained a deletion in the beginning. The Asn23 and Ser27 were however conserved in the other three sequences. The Glu24, also deleted in the sequence FLEX ROSE, was conserved in two sequences and contained substitution Glu24Gly in the sequence OLEI SOLI. The Ser88 was not directly conserved, but a possibly substituting Asp was present at the previous position.

Two of the other conserved positions, Trp21 and Gly26, were also deleted from the FLEX ROSE sequence. The hydrophobic positions Leu39, Leu73 and Ile104, were all conserved with the exception of the Leu73Cys substitution in the sequence MARI PURP. Conservative substitutions of Thr40Arg and Thr76Ser were prevalent. Thr90 was substituted

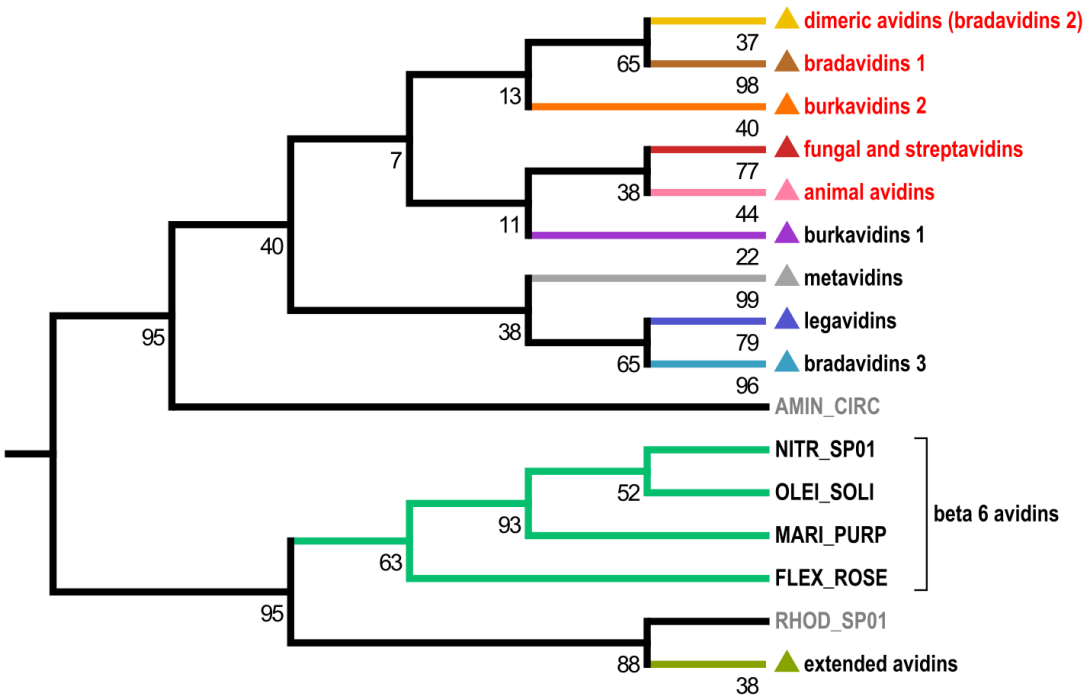


Figure 18. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the β6 avidins clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

with hydrophobic residues. The Ser93 and Gly94 residues were moved one position further. Lastly, the Gly126 residue was substituted non-conservatively to either Tyr or Ala.

Extended avidins

In the **extended avidins** clade, there was altogether 20 fully conserved amino acid residues (Fig. 19 and Fig. 23 for the phylogeny and MSA, respectively). Seven out of these were the biotin-binding residues, Asn23, Ser27, Tyr43, Ser45, Trp79, Trp92 and Asp128 (Table 4, Fig. 23). Further six were generally conserved among avidins: Trp21, Gly41, Gly58, Val77Ile, and Gly126. Although in the β8-strand, the distances between conserved residues, were longer (Table 4, Fig. 23). Additionally, there was a Gly at the location 98, possibly compensating for the missing Gly94. Other conserved positions included Gly19, Thr4848, Gly49, Gly52, Tyr54, Gln95, and finally a Leu at the position 130.

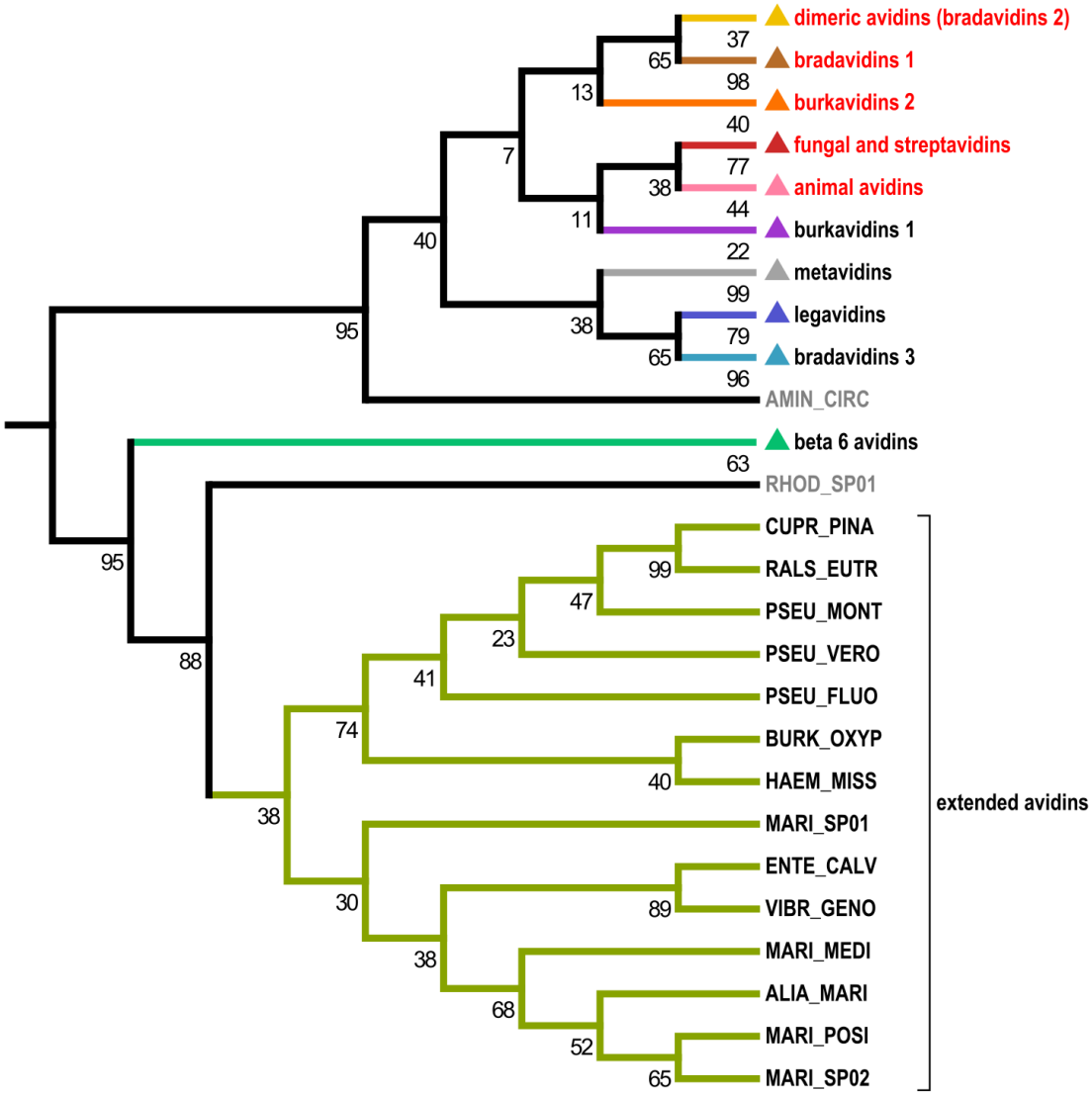


Figure 19. Phylogenetic cladogram tree of the verified and putative bacterial avidin sequences with the extended avidins clade expanded. Triangle marks the collapsed clade, red text the clades containing verified avidins, and grey text that the sequence was an outlier. The clades are coloured by identifiable subgroups as in Fig. 5. The tree was computed and visualized in MEGA6, and further visually edited in InkScape.

The biotin-binding positions contained homologues in most sequences with the exception of Trp108 that was substituted to non-aromatic hydrophobic residues. The Trp92 was shifted one amino acid further. Substitution of Leu110Ala was present as well as a substitution of Trp120Tyr or Trp 120Phe. The Phe residue occurred with a Pro residue, similar to the **dimeric avidins**, however aromatic residues aside from Trp have not been previously reported for this position.

Most of the generally conserved positions were also well preserved. Gly26 and Gly37 and the six hydrophobic positions, Val31, Leu39, Leu56, Leu73, Trp75 and Ile104, remained con-served. However, the Thr40 was not conserved, the Thr76 was substituted to a Ser or an Ala residue, Thr90 was changed to a Trp residue, Thr106 to Asn or His, and finally the Phe130 to a Tyr or Phe.

There were a few clear changes unique for this clade present, a hydro-phobic residue at position 17, a Met at 29, Asp or Asn at 87, and finally the β8-strand was elongated. The AAs between the Gly126 and Asp128 included a con-served polar residue followed by a con-served Tyr and another hydrophobic position. Similarly the AAs between Asp128 and Phe130 contained a Lys, Leu and non-conserved position.

Table 4. Residues present in the biotin-binding and conserved positions in each identified subgroup and fibropellins. The blue cells signify an amino acid residue seen in the AVD or Streptavd sequences, green cells signify an amino acid residue seen in other verified avidins, yellow cells signify a conservative change from the residues in verified avidins, and finally orange cells signify non-conservative change from residues present in the verified avidins. Fibropellin was included for comparison as it is verified not to bind biotin, but is considered to resemble avidins in terms of 3D-structure.

	Biotin-binding positions											
AVD/ strAVD	N12/ N23	Q24	S16/ S27	Y33/ Y43	T35/ S45	W70/ W79	S75/ S88	F79/ W92	W97/ W108	L99/ L110	W110/ W120	N118/ D128
Dimeric avidins	N	Q/E	S	F/Y	N	F/W	S/T	W	W	L	F/N/G/P	D
Bradavidins 1	N	Q	S	Y	N	F	S	W	W	L	W	D
Burkavidins 2	N/T	N/Q/S/E	A/S	F/Y	S/T	F/W	S/T/G	W	W/Y/D	F/W/Y/I/L	F/W	T/D
Fungal and streptavidins	N/S	Q/E	S	Y	S/T	F/W/Y	S	W	W	L	W	D
Animal avidins	N	K/D/E	S	Y	T	W	S	F/W	W	L	W	N/D
Burkavidins 1	N/S/G	R/D/E/G	S/G	F/Y	S/T	F/W	A/S	F/W	W	L/M	W	A/N/S/D
Metavidins	N/–	E/–	A/–	Y/C	shift	S/T	T/E	W	F/Y	A/V	R/N/G	A/T
Legavidins	N	Q/T/K/E	S	F	T	F/Y/L	S/shift	F/I/L/V	W/A/S/T	I/L/V	W/L/P	N/D
Bradavidins 3	N	Q/E	S	F	T	F/A/S/T/C	A/M/V/S/R	F/Y	W	V	W/WW	D
β6 avidins	N/–	E/G/–	S/–	Y	S	W	S/shift	W	Y	L	W	S
Extended avidins	N	A/S/D/E	S	Y	S	W	S/T/shift	W	N/H	I/L/M	F/Y/P	D
Fibropellin	N	E	D	Y	T	R	S	W	W	N	K	D

AVD/ strAVD	**Generally conserved amino acids** W10/ W21	G15/ G26	I20/ V31	G27/ G37	F29/ L39	T30/ T40	G31/ G41	L49/ L56	G51/ G58	F64/ L73	F66/ W75	T67/ T76
Dimeric avidins	W	G	F/I/L/V	Y/D/G	F/I/L/V	A/L/Q/S/T	G	A/L/V	G	I/L/V	F/L	Y/A/V/S/T
Bradavidins 1	W	G	I/V	G	I	A/S/T	G	A/M/V	G	I	F	S/T
Burkavidins 2	W	G	F/I/V	G	I/L/V	V/Q/S/T	V/G	Y/L/M	G	F/I/L/V	F	A/S/T
Fungal and streptavidins	W	N/Q/G	F/L/V	G/shift	F/I/L/V	V/S/T	G	L/M/V	G	I/L/V	F/W	A/V/T
Animal avidins	W	G	F/I/V	G/shift	F/M/V	T/R	G	I/L	G	F/V	F	S/T
Burkavidins 1	W	D/E/G	F/I/L	A/S/D/G/P	F/I/L/V	S/T	G	A/I/L/V/T	G	I/L	F	L/M/V/T/C
Metavidins	W/–	G/–	L/–	G	I/V	S	G	L	G	I	F	V
Legavidins	F/W/Y	G	L	N/S/D/G	I/L/V	S/T/K/E/–	G/–	I/V/C	G	F/L/M/V	F/I/L/N	S
Bradavidins 3	W	G	I/L	N/S/H/E/G	I/L/V	A/Q/S/R/E	G	I/L/V	G	A/I/L/V	F/V	A/V/N/S/T
β6 avidins	W/–	G/–	I	G	F/I	I/R	G	V	G	I/V/C	F	A/S/T
Extended avidins	W	H/G	I/L	N/Q/H/G	I/L/V	F/W/Y/A/I/ V/S/R/C	G	I/L/V/Q	G	A/I/L/V	F/I/L	A/S
Fibropellin	W	N	I	G	M	L	G	V	G	F	F	T

AVD/strAVD	V68/ V77	T77 /T90	T80 /S93	G81/ G94	L93/ I104	T95/ T106	G116/ G126	F120/ F130
Dimeric avidins	V/T	T	V/N/S/T	G	I/L/M	A/T	G	F
Bradavidins 1	V	T	T	G	F	V/G	G	F
Burkavidins 2	V	S/T	V/N/S/T	G	I/L	A/S/T	N/G	F/L
Fungal and streptavidins	I/V	A/S/T	S	G	I/L	A/T	G	F
Animal avidins	V	S/T	A/V/T	G	L	T	G	F
Burkavidins 1	V	S/T	A/V/S/T	G	I/L/M	A/S/T	N/G	F/W
Metavidins	T	T	T	G	I	A/L	E/G	F
Legavidins	F/I/V/N	I/L	I/S/T	G	I/L	A/T	T/G	F/Y
Bradavidins 3	–	A/V/C	A/S/T	G	I/L/M	A/L/M/T	N/S	F
β6 avidins	V	W/A	shift	A/G	F/I	T	Y/A	F
Extended avidins	I	W/V/S	A/V/S/T	S/G	L/M	L/M/V	shift	F/Y
Fibropellin	V	T	T	G	L	T	G	W

AA occurs in streptavidin

AA occurs in a verified avidin

conservative substitution

non-conservative substitution

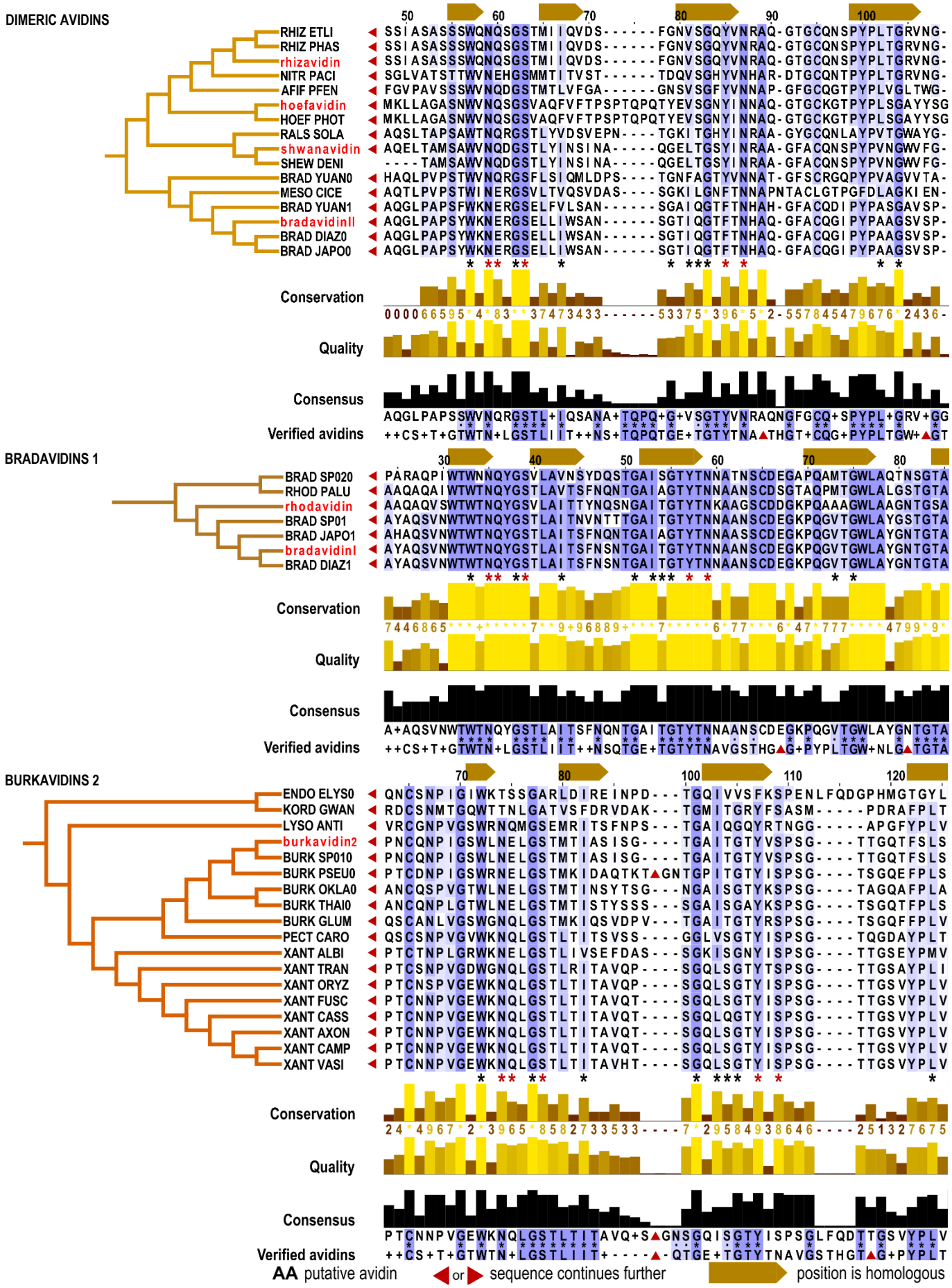
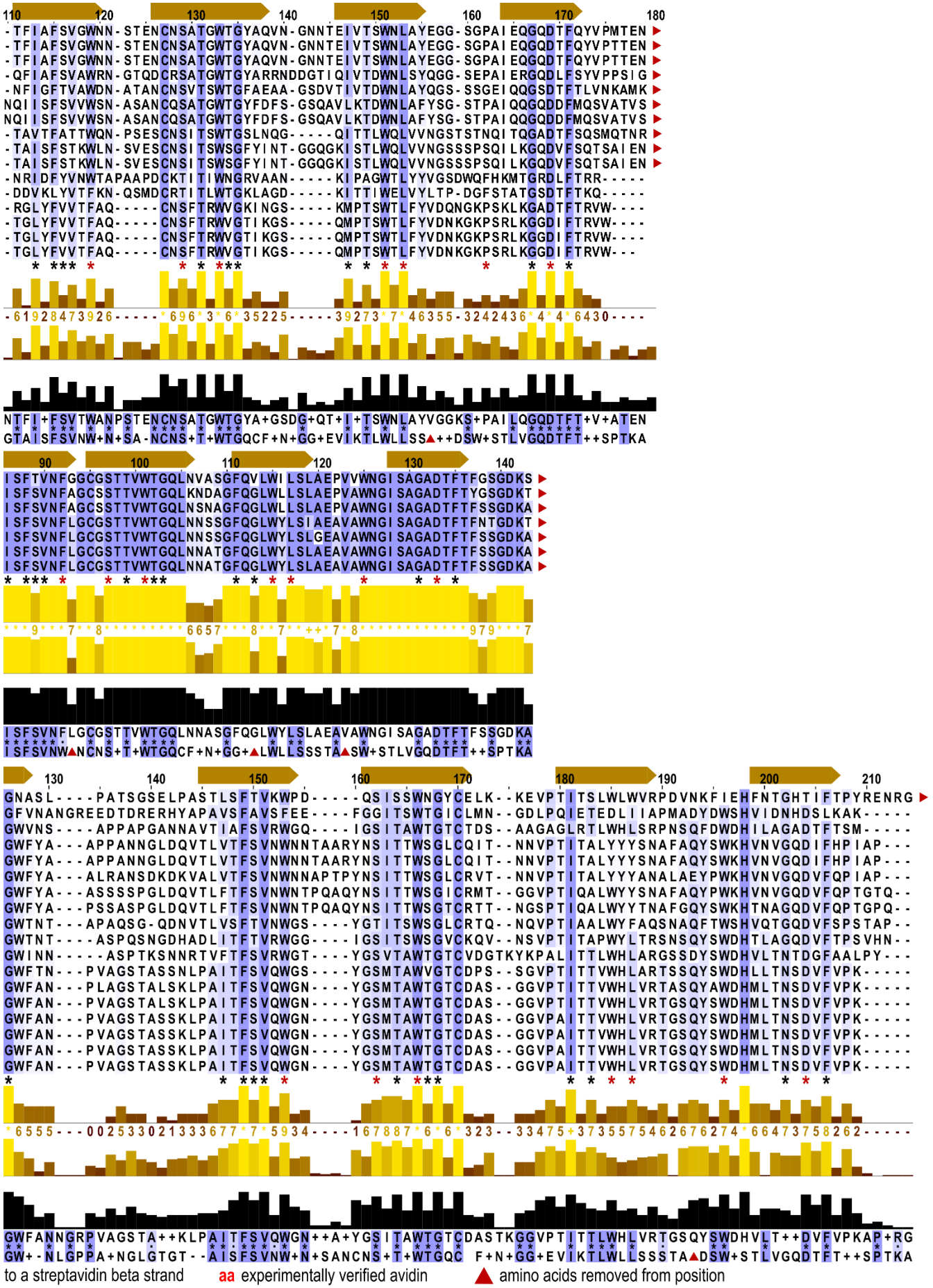


Figure 20. Multiple sequence alignments of the subgroups dimeric avidins, bradavidins 1, and burkavidins 2. Red triangles in the alignment mark sequence continuation; red triangles in the consensus sequence line mark that amino acids have been removed from the position; the brown arrows above the sequence mark the locations homologous to streptavidin β-strands; asterisks mark the biotin-binding (red) and the generally conserved positions (black). Sequence names are coloured red, if the sequence is a verified avidin.


to a streptavidin beta strand
aa experimentally verified avidin
amino acids removed from position

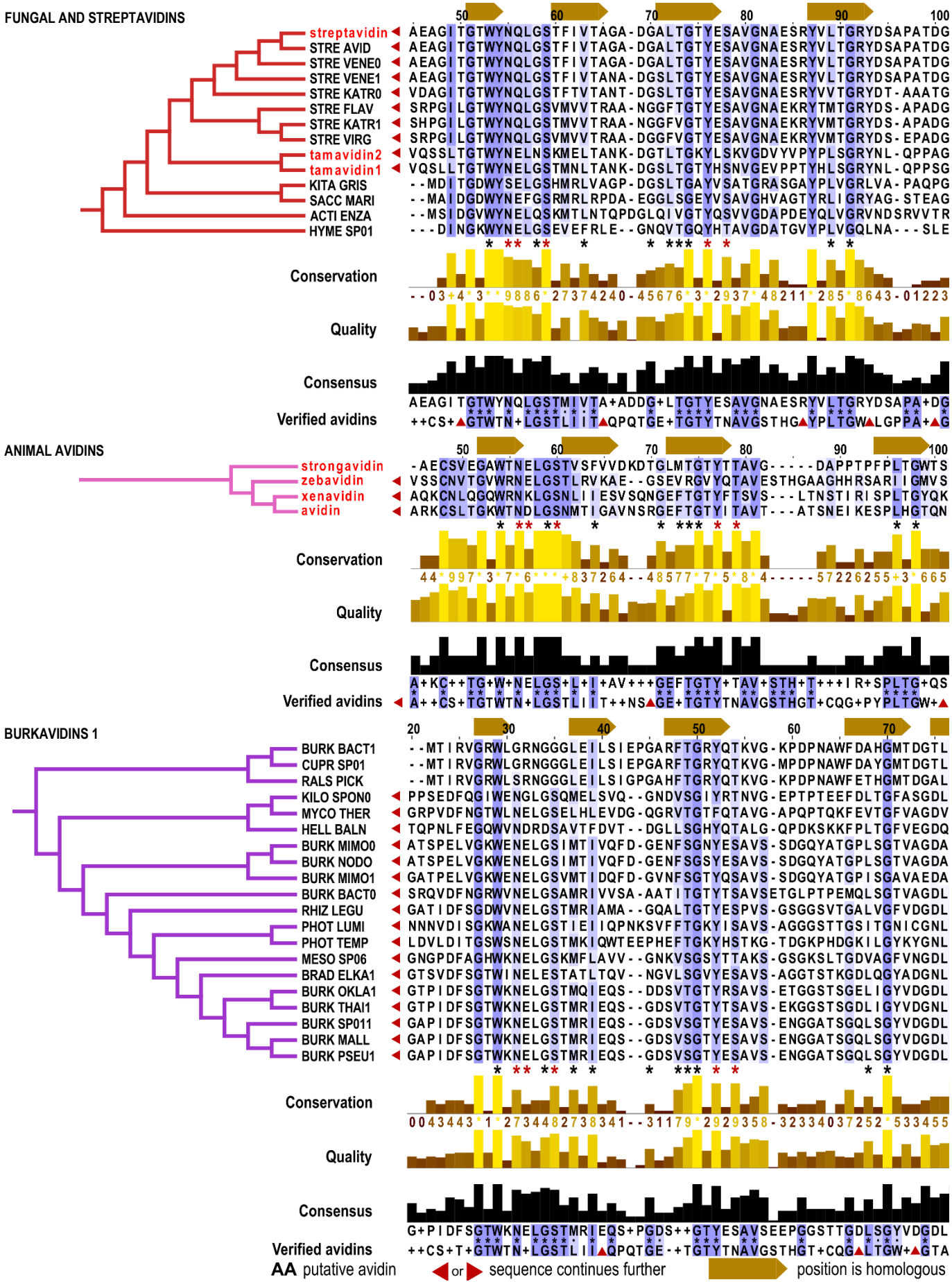
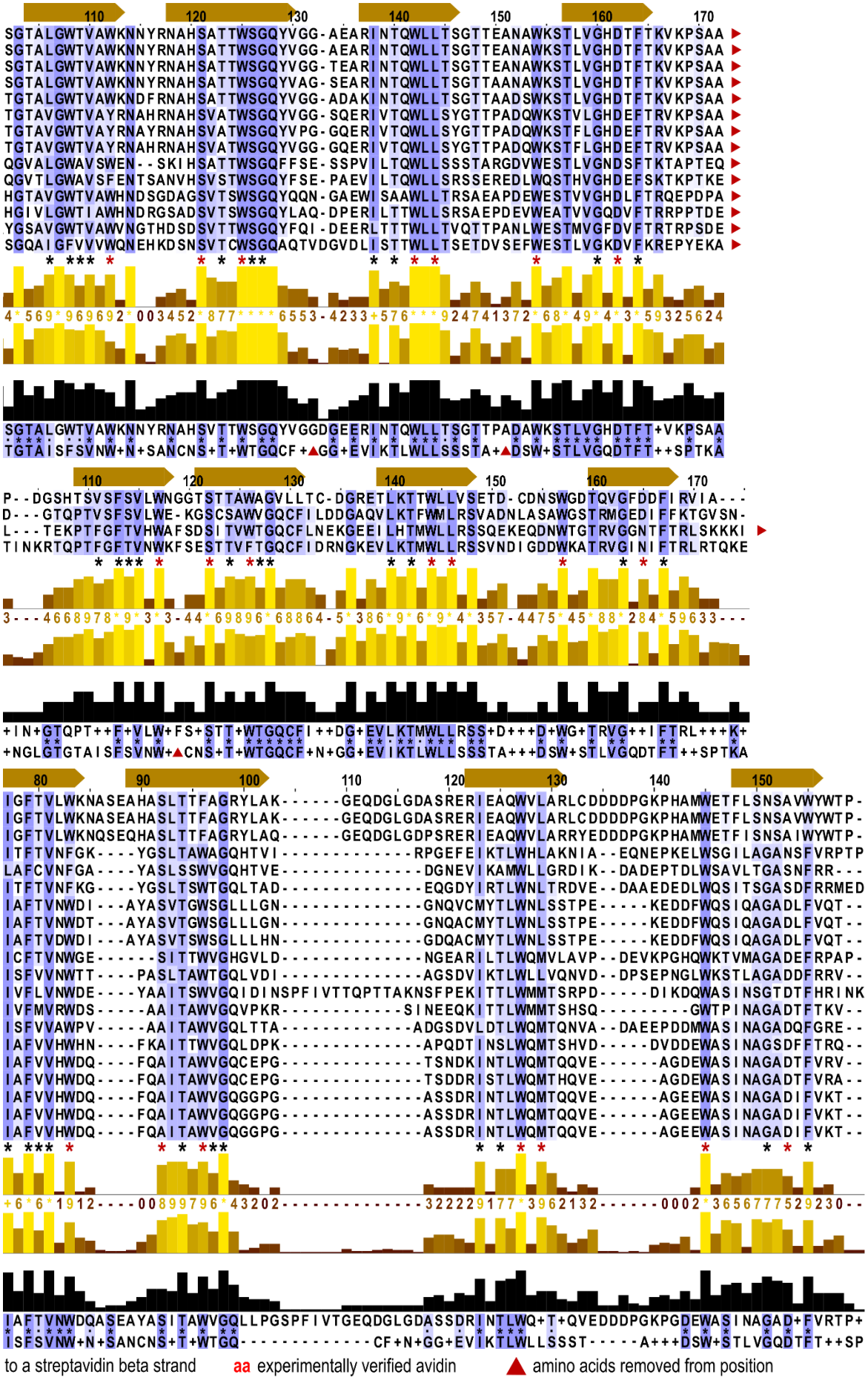


Figure 21. Multiple sequence alignments of the subgroups fungal and streptavidins, animal avidins, and burkavidins 1. Red triangles in the alignment mark sequence continuation; red triangles in the consensus sequence line mark that amino acids have been removed from the position; the brown arrows above the sequence mark the locations homologous to streptavidin β-strands; asterisks mark the biotin-binding (red) and the generally conserved positions (black). Sequence names are coloured red, if the sequence is a verified avidin.


to a streptavidin beta strand
aa experimentally verified avidin
amino acids removed from position

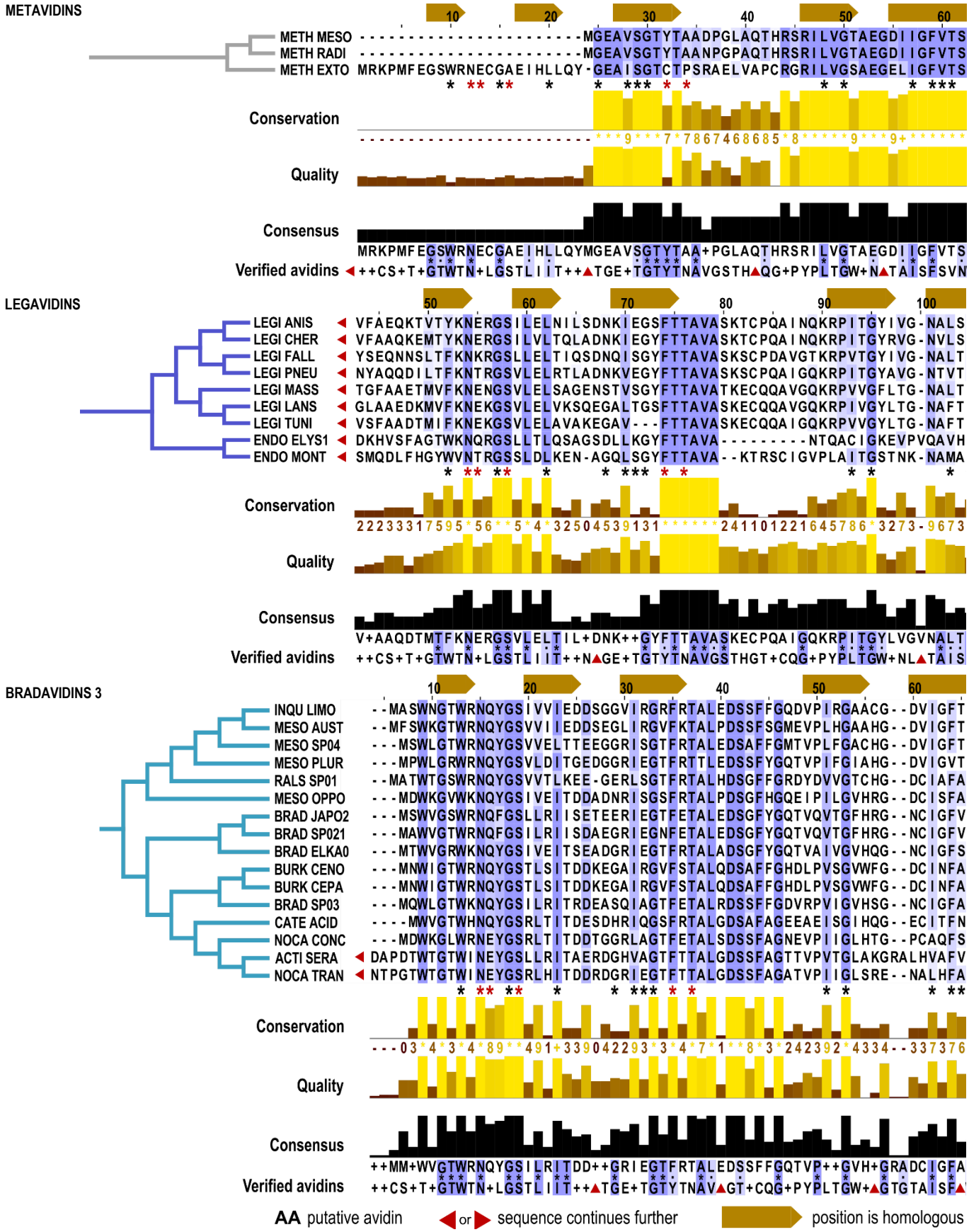
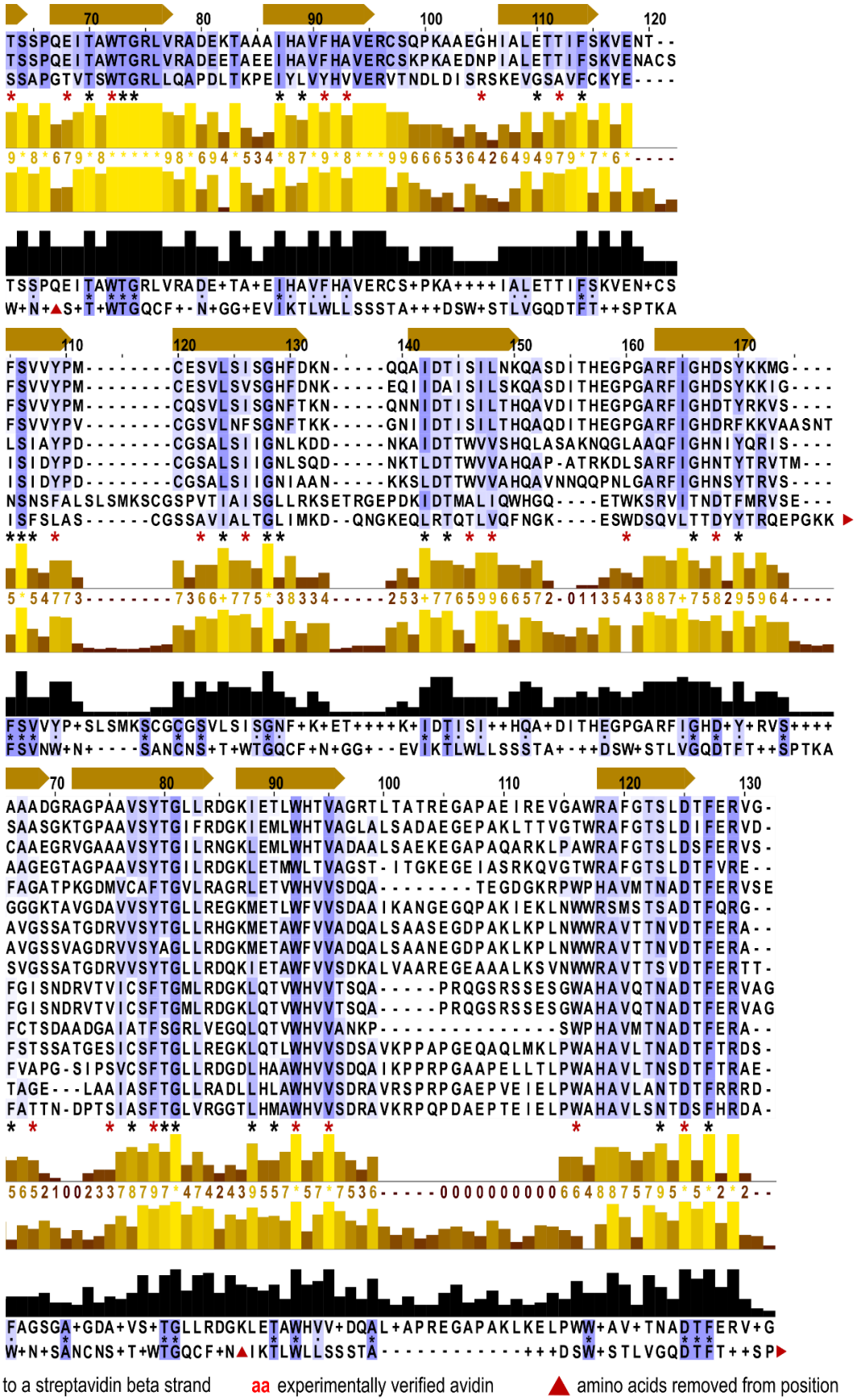


Figure 22. Multiple sequence alignments of the subgroups metavidins, legavidins, and bradavidins 3. Red triangles in the alignment mark sequence continuation; red triangles in the consensus sequence line mark that amino acids have been removed from the position; the brown arrows above the sequence mark the locations homologous to streptavidin β-strands; asterisks mark the biotin-binding (red) and the generally conserved positions (black). Sequence names are coloured red, if the sequence is a verified avidin.


to a streptavidin beta strand
aa experimentally verified avidin
amino acids removed from position

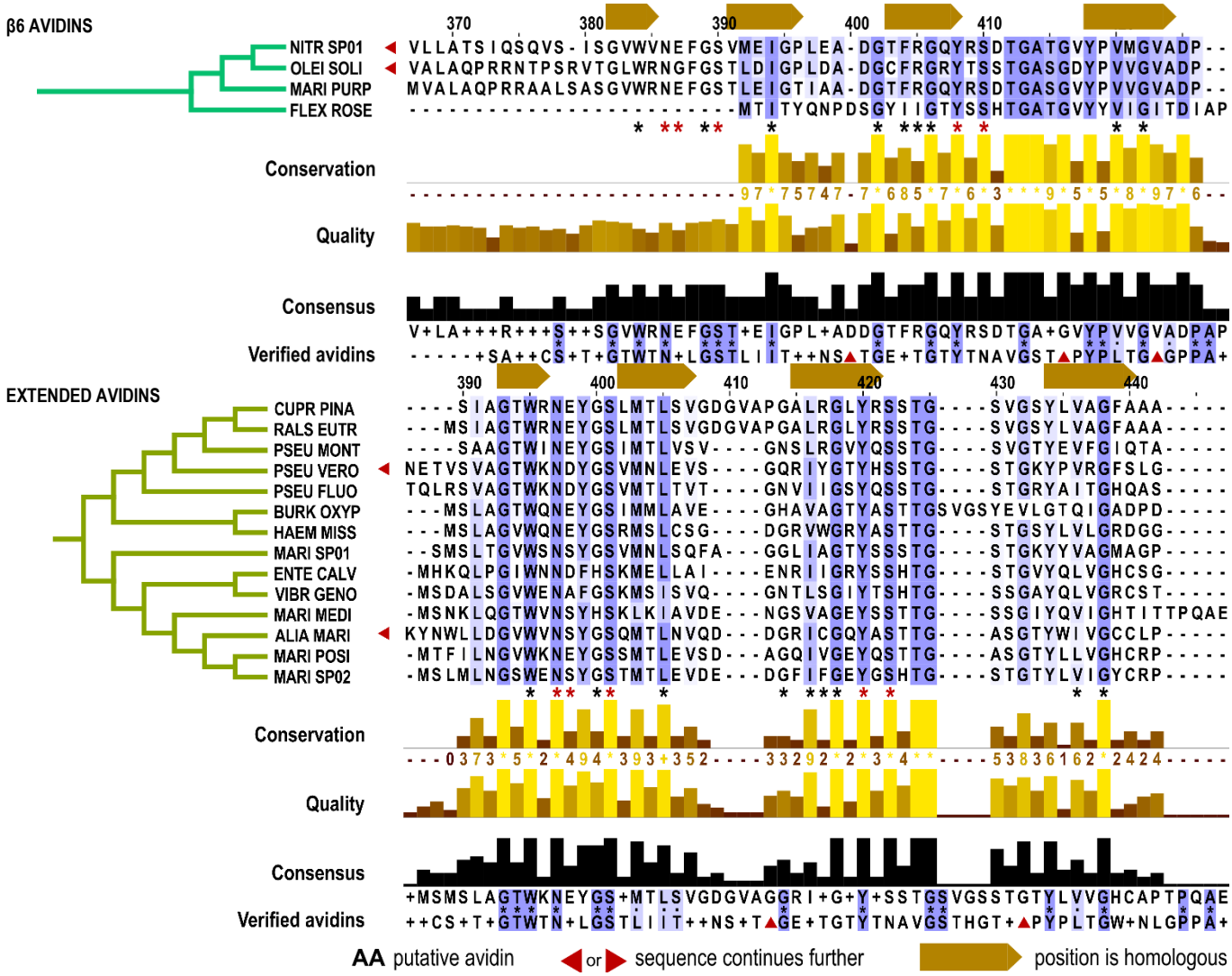
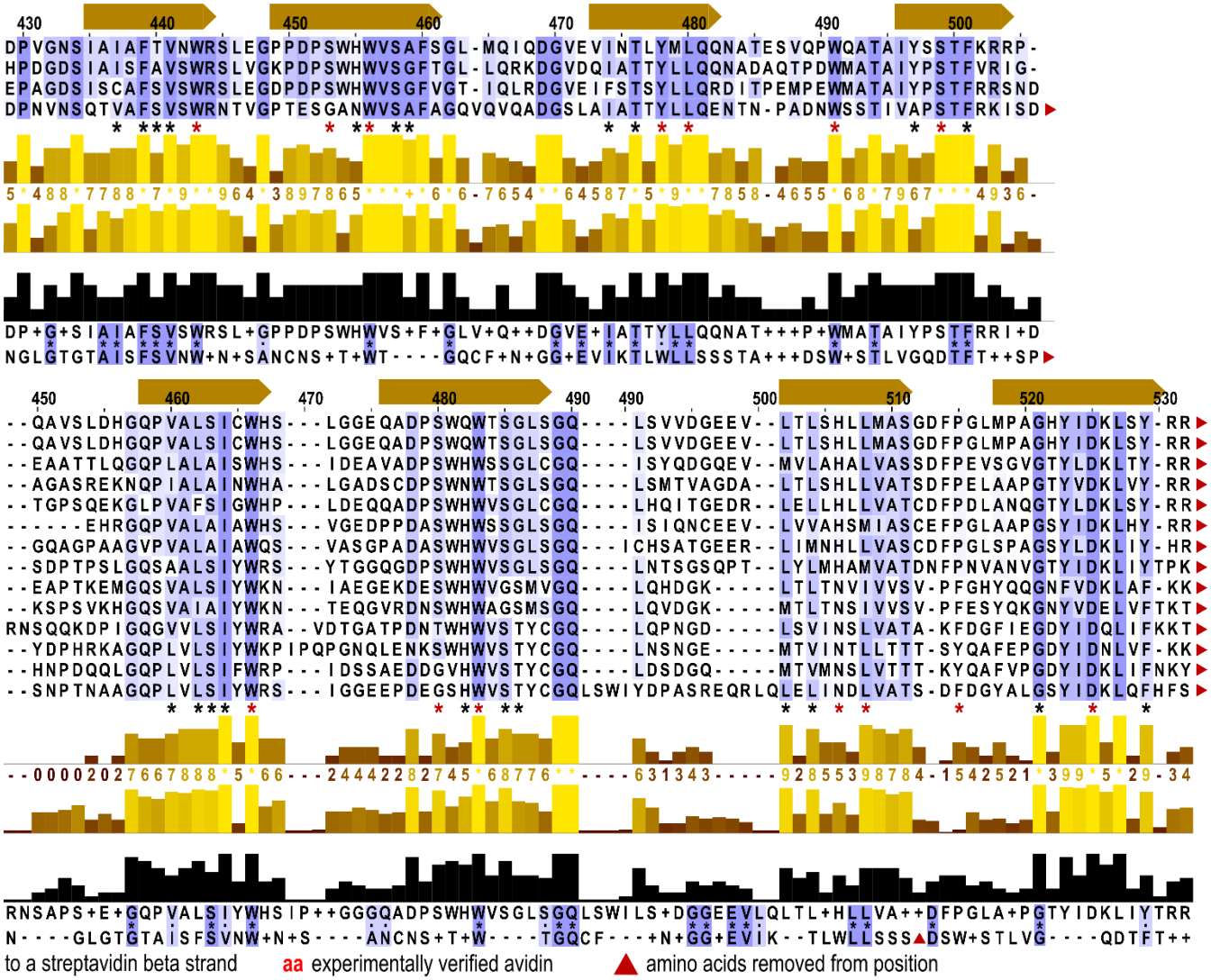


Figure 23. Multiple sequence alignments of the subgroups β6 avidins, and extended avidins. Red triangles in the alignment mark sequence continuation; red triangles in the consensus sequence line mark that amino acids have been removed from the position; the brown arrows above the sequence mark the locations homologous to streptavidin β-strands; asterisks mark the biotin-binding (red) and the generally conserved positions (black). Sequence names are coloured red, if the sequence is a verified avidin.


to a streptavidin beta strand
aa experimentally verified avidin
amino acids removed from position

5.4 Enrichment analysis

90 Gene Ontology -terms (GO-terms) were found to correspond to the genomic features in the proximity of avidin. The terms were tested with Fischer exact test for the probability of them associating generally with avidin genes. The full table of the results can be seen in Table Appendix C. 27 terms were enriched with p-value less than 0.05 and a total of 38 terms were enriched with p-value less than 0.1.

The list was strikingly dominated by GO-terms related to DNA-processing and integration of mobile elements (light grey Table Appendix C). Altogether there were 19 terms related to these functions and for 14 of these the enrichment p-value reached below 0.1. Another highly represented group of individual GO-terms was a set of 11 GO-terms all related to the most fundamental metabolic pathways including fatty acid metabolism, polysaccharide metabolism and tricarboxylic cycle (medium grey Table Appendix C). Additionally, 7 of these terms had p-value reaching below 0.1.

In the list, there were only three terms directly related to defence pathways: GO:0051607 (defence response to virus), GO:0005976 (virion) and GO:0006950 (response to stress) (dark grey Table Appendix C). While the two first of these terms were enriched with p-value less than 0.05, the last one was enriched with a modest p-value of 0.171.

6 Discussion

The study began with a set of verified avidins that were used for BLAST querying putative avidin sequences from bacteria in the public databases. For protein sequences RefSeq, PDB, GenBank, and UniProtKB databases were used, while for nucleotide sequences GenBank, EMBL Nucleotide Sequence Database, and DDBJ databases were queried. A new avidin focused database, DATAvidin, was constructed from these putative sequences. Structural MSA was built using the verified avidin sequences and a set of 113 unique sequences from the putative avidins were selected and aligned against the structural alignment. Furthermore phylogenetic cladogram trees were created using the MSA. Lastly, ten origin species' genomes of the verified and putative avidin sequences were chosen for an enrichment analysis. The genomes were picked across the identified avidin sub groups in the phylogenetic tree to bring variety in the material set. The enrichment analysis assessed the avidin genes' association with other genes related to specific cellular functions.

6.1 DATAvidin

The goal for the new database, DATAvidin (http://86.50.169.79:3000), was to collect information of both previously verified and new putative avidins in one place together with the information of the key AA positions. This was aimed to ease the identification of new interesting targets for experimental verification, but also to help evaluating the effects of AA changes in avidin sequences or the engineering of AA changes with desired effects.

DATAvidin database, indeed offers an interface to selectively browse a collected set of putative avidin sequences or to BLAST query this set with a sequence of interest. This, together with the option to build quick alignments against a structural MSA of verified avidins, can streamline searching for and deciding upon interesting experimental targets. The possibility to inspect sequence features, at given position of the constructed alignment, is aimed to help evaluating potential effects of AA changes in in the given position.

In addition to the above functions, two more were planned yet not implemented due to the restricted time. The alignment view was proposed to include an embedded 3D-view of the Streptavd's structure, which would highlight the chosen amino acid position. The 3D-view would have made the evaluation of positions and AA changes more intuitive. The other function that was not implemented, but was suggested and planned was an interactive phylogenetic view to select sequences from. The chosen sequences or sequence sets could be used for BLAST

query and/or alignment. This function could have brought the bacterial species and sequence clustering in as another element, when deciding upon interesting experimental targets.

Aside from these unimplemented functions, a few other factors, limit the usefulness of the database currently. Firstly, the database includes only bacterial sequences and uses Streptavd as the reference sequence for feature positions. While DATAvidin can be expanded in future to cover also eukaryotic species, it is recommendable to use a eukaryotic avidin, such as AVD, as a reference sequence for these. Furthermore, the eukaryotic exon–intron structures have to be accounted for and adding another reference sequence for eukaryotic DNA entries should be considered. Secondly, the sequence feature entries in the database do not yet include information on the effects of AA changes in that specific position. Entries are designed to contain this information, however, the addition of this data was planned to work by user-submission basis, which has not yet been implemented.

6.2 MSA and phylogeny

Bootstrapping (BTSP) is a commonly used method to evaluate the tree credibility in phylogeny (Felsenstein 1985; Efron et al. 1996). In the phylogenetic context, the method constructs new trees based on permutations of the underlying alignment and returns the amount of trees supporting the clade in question as percentage. Generally, BTSP values above 70 are considered to give a strong support for the node, while the support from values below 50 is not considered significant (Efron et al. 1996).

The BTSP values across the phylogenetic cladogram tree of the putative avidins were greater than 70 for 50% of the nodes, while only 30% of the nodes had their BTSP value smaller or equal to 50, as shown in Table 3. For comparison, the Table 3 also presents these percentages for each individual major branch, as well as for the nodes between the major branches. While not excellent, the values were satisfactory considering the specifics of the data. The low sequence identities of avidins make the construction of good quality MSAs challenging for them. Furthermore, the sequence set was large and contained multiple sites prone for insertions and deletions, both of which complicate the construction of reliable MSA. For external control, the topology of the putative avidin set was compared to that of the verified avidin set. Both cladograms followed closely the same topology increasing the credibility of the trees.

6.2.1 Bacterial heterogeneity among the verified avidin groups

Interestingly, in the putative avidins' tree, the branches containing verified avidin leaves clustered together into a super branch (Fig. 8). For convenience, this branch was named as

verified avidins. This superclade included the clades named as **dimeric avidins**, **bradavidins 2**, **burkavidins 2**, **fungal and streptavidins**, **animal avidins**, and **burkavidins 1**. While the **burkavidins 1** does not contain an experimentally verified avidin, it contains two sequences that have previously been suggested as functional avidins, Burkavd 1 and Xantavd (Helppolainen et al. 2008; Sardo et al. 2011). The putative avidin sequences within this superclade were dispersed into the branches with the verified sequences rather than clustered together on their own. Also, the BTSP values were generally higher within each child branch of **the verified avidins** than between branches (Table 3). This suggests that these putative avidin sequences within this branch are true members of the avidin family.

Although, most of these putative avidin sequences in the **verified avidins** branch originated from species closely related to the bacteria known to contain avidin gene, some were from unexpected species not previously reported to contain such gene (Fig. 8). Thus it appears the avidin family is more widely spread within prokaryotic kingdom than previously thought. Among the new species were bacteria isolated from desert; dry and wet soil; marine, lake and brackish water environments (Fig. 8, Table Appendix A). Furthermore, the species had complex and diverse relationships with other organisms: several species were involved in skin or endogastric microflora in animals, some were isolated from plant microflora, some were capable of opportunistic pathogenicity while others acted as symbiotic companions for the host, and some species secreted antibiotic or antifungal proteins. The bacterial source organisms and their habitats are catalogued in the table of appendix A with references.

At least three of the species were reported to have the ability to thrive in extreme or unusual conditions, one species was isolated from contaminated soil, another from deep sea black smoker and one showed UV-resistant qualities (Fig. 8, Table Appendix A). Avidin seems to be present often in species with strongly adaptive nature. Another interesting detail is the heterogeneity of the bacterial habitats even within the branches. Perhaps the most striking example of this was set by the **dimeric avidins**. This branch contains sequences from the root nodular symbiotic species, a plant pathogen, marine, and sediment bacteria, as well as, a species isolated from sponge surface flora (Fig. 8, Table Appendix A). How or why bacteria from such diverse habitats express a similar version of the avidin will be discussed later.

It is important to note that bacterial classification is not as straightforward as with other organisms and the species are subject to change as the bacterial genomes are studied further. Some of the sequences in this study, as well, have had their origin species reclassified either before or during the study. While the species names were updated in the Table appendix A

whenever new names were indicated in the literature or databases, several of the species names may yet be outdated or change in the future.

6.2.2 Bacterial heterogeneity among the putative avidin groups

Five branches were left outside of the **verified avidins** branch: **metavidins**, **legavidins**, **bradavidins 3**, **β6 avidins**, and **extended avidins**. Out of these branches only **metavidins** seemed immediately unlikely to contain any functional avidins. The branch consisted of only three sequences; two of these sequences were truncated at the N-terminus thus missing the β1 and β2-strands; and finally the biotin-binding positions were poorly conserved, as the branch showed more non-conservative substitutions in these, positions than conservation.

However, the last four branches are especially interesting. They contain more conservation in the biotin-binding positions than non-conservative substitutions, but do not branch together with any of the known and verified avidins. **Legavidins** origin species consisted almost exclusively of plant or human pathogenic *Legionella* genus species able to also live as amoebae intracellular parasites (Fig. 8, Table Appendix A). The last two species, *Endozoicomonas elysicola* and *Endozoicomonas Montiporae*, are present in the sea slug or coral microflora instead. It is notable too, that *Legionella* are aquatic bacteria as well (Fig. 8, Table Appendix A). However, the sequences from these two species contained a more diverse sequence and clustered as outgroup within the **legavidins** branch.

Bradavidins 3 branch was named for the multiple Bradyrhizobium species that dominated the clade. At first, it seemed that the group also contained a previously uncharacterized protein sequence from same species the Bradavd I and Bradavd II were isolated from. However, the BRAD JAPO2 sequence in **bradavidins 3** is isolated from a different strain of *Bradyrhizobium japonicum* than the previously characterized avidins (Fig. 8, Table Appendix A). The strain *B. japonicum USDA110*, expressing the Bradavd I and Bradavd II, was recently reclassified as a member of *B. diazoefficiens* and thus the sequences BRAD JAPO0 and BRAD JAPO1 from the **bradavidins 1** and **dimeric avidins** branches, respectively, are not from the same species as the BRAD JAPO2 within the **bradavidins 3** branch (Delamuta et al. 2013). However, the sequence BRAD SP021 is from the same unnamed species, *Bradyrhizobium sp. WSM3983*, as the BRAD SP020 in **bradavidins 1** branch (Fig. 8, Table Appendix A). Thus there might yet be some overlap in the presence and spread of this new putative avidin group and **bradavidins 1**. This connection already lends support to the **bradavidins 3** as true members of avidin family and makes the group specifically interesting. Overall, the sequences were predominantly from the root nodular symbiotic bacteria, either

belonging to the aforementioned *Bradyrhizobium* genus or to the *Mesorhizobium* genus (Fig. 8, Table Appendix A). The other species of origin included human and plant pathogenic, decomposing, and human skin microflora bacteria (Fig. 8, Table Appendix A). The group is rather similar to the **verified avidins** in this aspect.

The last groups, **β6 avidins** and **extended avidins**, are more complex to evaluate. While they showed similar amount of conservation in biotin-binding positions compared to verified avidins, their sequences contained significant larger scale changes that will be discussed later. Aside from the sequence peculiarities, the branches were rather heterogeneous in the sequence origins and had rather low BTSP values compared to the rest of the branches (Table 3).

β6 avidins was a rather small branch, containing only 4 sequences, none of which were from closely related species. One of the sequences was form a soil bacteria able to degrade contaminants, while the rest three were aquatic (Fig. 8, Table Appendix A). The aquatic species included an alkaliphile, photosynthesizing and fish pathogen species. However, it is difficult to draw any conclusions since the group is so small. The apparent heterogeneity or the dominance of aquatic species might be simply a result of bias.

The **extended avidins** branch contains two sub branches that could be considered their own separate groups. However, since these new groups would have been rather heterogenic in nature as well as small, they were inspected together. There is some distinction in the origin species of the sequences between these sub branches: the second, clockwise in the Figure 8, contains only marine species, while the first included species that primarily habit soil. Some of these soil bacteria could be considered extremophiles; one was a metal resistant species isolated from volcanic sludge, another was acidophilic, and third involved in contaminant degradation (Fig. 8, Table Appendix A). Yet the three *Pseudomonas* species in this sub branch appear all to be sub species of the same bacteria with complex endosymbiotic and endoparasitic relationship with soil amoebae (Fig. 8, Table Appendix A). Similarly the marine species sub branch of the **extended avidins** contained sequences mostly from bacteria with complicated dependencies with higher organisms. The origin bacteria included microflora participating species found in seaweed and salt marsh grass, antibiotic secreting bacterioplankton, and a possibly obligate symbiont of other bacterial species able to degrade contaminants (Fig. 8, Table Appendix A).

Finally, these complex interplays with both higher organisms and other prokaryotes seem to be the only repeating pattern among both the verified and putative avidin producing bacteria. While, this supports the hypothesized role of avidin as antimicrobial agent, it could also be bias in the material. The bacterial species that live in complex environments are also likelier to be specific targets of interest for humans. Such bacteria are more likely pathogens,

affect the agriculture either positively or negatively, degrade pollutants, survive in extreme environments or produce antibiotics. Thus these types of bacteria might be more often characterized and their genomes and proteomes are more often available in the databases. This bias considered, it cannot be excluded that avidin could appear in some more simple and solitary bacteria as well.

As the avidins evidently are more varied both in sequence and presence over the prokaryotes, it is to be expected that the amount of verified family members will grow as new bacterial species genomes and proteomes are sequenced. This will both ease and complicate the research of new avidins. The more avidin sequences are verified, the more effectively new putative members can be identified. Yet, if the putative avidins identified even in this study are verified as truly functional members of the family, the avidin family's signature features and known conservative positions need to be reconsidered. The branchwise alignments and the inspection of the biotin-binding and positions previously considered conservative are showing more substitutions, including non-conservative ones, than expected. Instead of a steady sequence signature, the family seems to favour a set of positions from which almost a third can be radically changed. Considering the sequence identities between the known avidin family members are as low as 20 %, this might not be a surprise. However, if the identity across the avidin family proves to be even lower, there is no effective methods to reliably identify new avidin family members that lie undiscovered beneath that identity threshold.

6.2.3 Amino acid changes

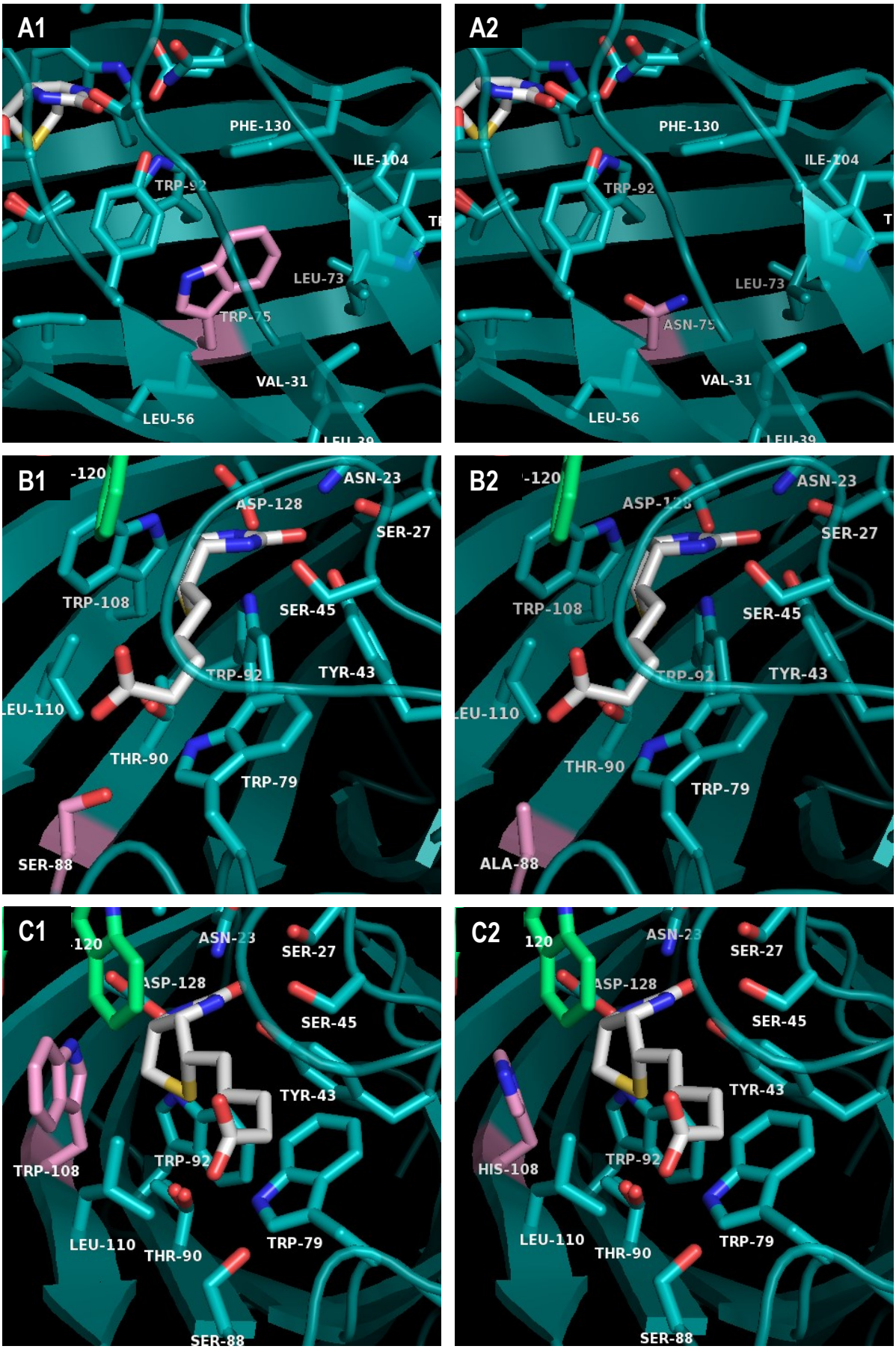
Most of the avidin clades contained some changes unique for their sequences. Some groups also featured changes or sets of changes similar to each other. Interestingly, in this kind of wider examination of the avidin family sequences, some of the positions previously considered conserved showed lot of variation. For four conserved positions that showed previously unreported substitutions were illustrated (Fig. 24).

Especially the **dimeric avidins** group was interesting to compare with the rest of the avidin clades. The generally conserved AA Gly19 was substituted to a polar Ser residue in the **dimeric avidins**, a change unique for these sequences. Further polar substitutions were present at positions 25, 109 and 125. In all other verified avidin sequences, these positions have remained hydrophobic. Together with the change of Trp120Pro these substitutions are known to contribute to the avidin dimericity and thus the comparisons to other clades in these positions can lend a clue of the possible unusual oligomerization in the putative avidin sequences (Laitinen et al. 2006; Helppolainen et al. 2007; Meir et al. 2009).

Similar to these AA changes, the **legavidins** clade position 25 contains a non-conservative substitution. However, this clade's sequences feature AAs with stronger polarization or even a charge under the right conditions: an Arg or a Lys. Furthermore, the **extended avidins** and some of the **burkavidins 1** clade sequences contain a polar residue at the position 109. The Val125 position contains some polar substitutions in half of the putative sequences in the **burkavidins 2** clade and in most of the **bradavidins 3** clade sequences. Interestingly, the position presumably homologous to the Val125 in **extended avidins**, shows no conservation at all and presents an array of hydrophobic, polar and even charged residues. Finally, the substitution of Trp120Pro was present in most of the **extended avidins** and some **legavidins**. In the **extended avidins** this substitution was accompanied with a Phe either directly before or after the position. In rest of the **legavidins** sequences a substitution of Trp120Leu was present. In mutational studies, the Trp120Ala substitution decreased the Streptavd affinity to biotin by third (Laitinen et al. 2006). Yet, it is possible the bulkier Leu residue does not impair biotin-binding as significantly, as an Ala would at this position. Altogether the clades **burkavidins 1**, **legavidins**, **bradavidins 3**, and **extended avidins** shared some specific changes with the **dimeric avidins**. Therefore, they could be hypothesized to show some instability in their tetramer oligomerization. However, sequences from each clade should be verified to produce a functional avidin protein and then evaluated further experimentally.

In the beginning of the sequences, in β1 to β3-strands, most of the unusual changes occurred in the **legavidins** clade. Aside from the similarities with the **dimeric avidins**, the **legavidins** also featured the substitutions of Gly19 to aromatic residues as in the **bradavidins 1**. Additionally, the usually polar position 42 contained an aromatic, albeit also polar, Tyr residue and the following Tyr43 was substituted with a Phe residue. The change Tyr43Phe was present also in **bradavidins 3** and has been shown to increase the K_d of Streptavd from biotin by eight-fold (Laitinen et al. 2006). However, while this lead to decreased affinity to biotin in the Streptavd, the effect in these new sequences can only be assessed experimentally.

The β4-strand is rather variable and did not contain any clearly prevalent changes in any clades. However, from the β5-strand onwards the amount of changes seems to rise. Especially the **bradavidins 3** clade contained noteworthy AA changes. The Trp79 and the Ser88 were not conserved at all in this clade and the Thr90 was substituted to an aromatic Phe residue. Mutation of Thr90Ala has been linked to increase of the K_d-value in the Streptavd (Laitinen et al. 2006). This region contained also the changes of Trp75Leu in the **extended avidins** clade, and Trp79 to TyrPro or ProTyr in the **legavidins** clade.


A1
PHE-130
ILE-104
LEU-73
VAL-31
LEU-56
A2
PHE-130
TRP-92
LEU-73
ASN-75
VAL-31
LEU-56
B1
ASN-23
ASP-128
SER-27
TRP-108
SER-45
TYR-43
THR-90
TRP-79
SER-88
B2
ASN-23
SER-27
TRP-108
SER-45
TYR-43
THR-90
TRP-79
ALA-88
C1
SER-27
SER-45
TYR-43
TRP-108
TRP-92
LEU-110
THR-90
TRP-79
SER-88
C2
SER-27
SER-45
TYR-43
HIS-108
TRP-92
LEU-110
THR-90
TRP-79
SER-88

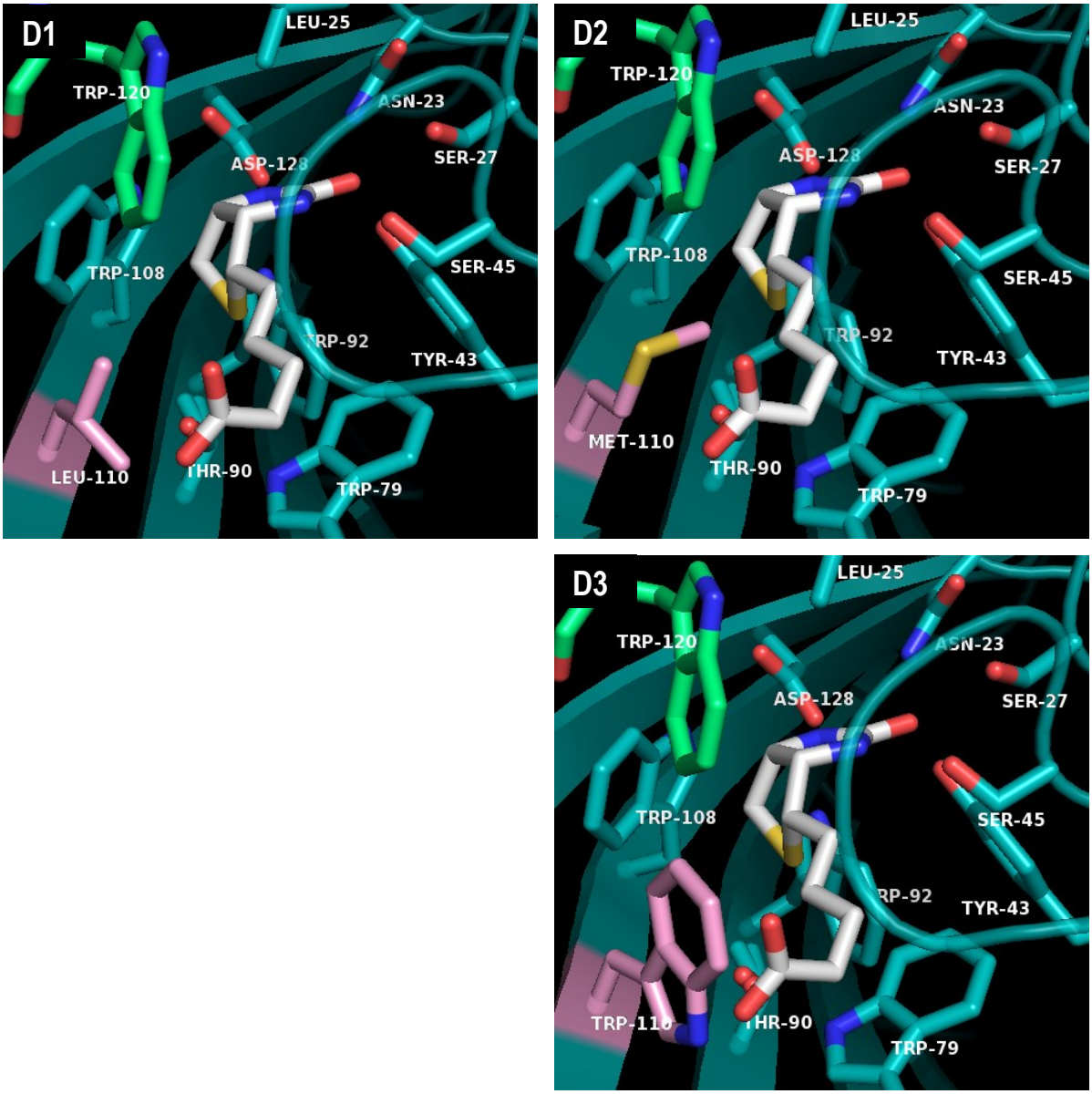


Figure 24. The biotin-binding site of the wild type streptavidin, as well as streptavidin with one amino acid substitutions. The panels A1 and A2 depict the Trp75 and its substitution to an Asn residue, respectively; the panels B1 and B2 depict the Ser88 and the substitution of Ser88Ala, respectively; the panels C1 and C2 depict the Trp108 and the substitution of Trp108His; finally, the panels D1, D2, and D3 illustrate the Leu110 and the substitutions of Leu110Met and Leu110Trp. The streptavidin ribbon model of the biotin-binding site and side chains are illustrated in teal, the Trp120 residue borrowed from the neighbouring subunit in green, the biotin in light grey and the amino acid of interest in pink. The oxygen atoms are shown in red, nitrogen in blue and sulphur in yellow. The used streptavidin structure was 3RY1 (Protein Data Bank) and the substitutions were modelled and visualized with PyMOL.

Further in the sequence, the biotin-binding Trp108 was changed to polar residues in the **legavidins** and the **extended avidins**. The **extended avidins** featured among other changes, also a Leu110Met substitution in two sequences. The special effects of this change to the biotin-binding site are illustrated in the Figure 24 panels D1–D2. Also the highly conserved Gly126 position featured a polar residue in some of the **burkavidins 2** and all of the **bradavidins 3** clade sequences. Finally the Phe130 position contained a change to a Tyr residue in the **legavidins** and **extended avidins**, while in the **burkavidins 1** clade the position featured Ile or Leu residues. Several other changes were present in the **extended avidins** and the **β6 avidins** clades, but they will be considered separately later.

As the **metavidins** showed more non-conservative changes than the other clades, it is briefly considered here on its own. Aside from two of the three sequences missing the first two β-strands, the branch contained three non-substitutions of the biotin-binding positions Ser27Ala, Ser45Ala and Trp79Thr. Additional, tree conservative substitutions were present at the biotin-binding positions: Ser88Glu, Trp108Phe, and Asp128Thr. Furthermore, the Val77 was changed to a polar Thr, Thr90 to a hydrophobic Met, and Gly126 to polar residues. All these changes considered, the clade does seem interesting as it either does not contain functionally biotin-binding avidins or it represents completely new type of avidin.

Aside from the avidins in the **fungal and streptavidins** clade, all of the verified avidins have contained two or more Cys residues in their sequence and only one of the putative avidins clades, **β6 avidins**, did not contain Cys residues either. The Cys pairs in avidins have been shown to form C–C bridges within the subunit and to stabilize its β-barrel structure (Rosano et al. 1999). Two clearly different sets Cys positions could be seen in the clades. The **dimeric avidins** clade and the closely related **bradavidins 1** clade contained conserved Cys residues in the loop 3–4 and the beginning of β6-strand. Again, the **legavidins** sequences showed similarity to the **dimeric avidins**, as the clade contains well conserved Cys residues in the corresponding positions.

The **animal avidins** and the **burkavidins 2** formed the other group with similar Cys positions. In these sequences the first Cys resides at the very beginning of the sequence, position 14 or 15, and the second Cys is present in the end of the β6-strand. None of the putative avidins clades matched this pattern, although several individual sequences contained one Cys residue in either of these two positions. Instead, the rest of the Cys positions in the putative avidins clades, **burkavidins 1**, **metavidins**, **bradavidins 3**, **β6 avidins**, and **extended avidins**, were rather scattered in the clades and several sequences contained either only one or no Cys at all.

In the end, all of the putative avidin clades are candidates for further experimental verification, since they feature each changes that are previously not observed among avidins. **Burkavidins 1** clade contains two previously reported, but experimentally not verified sequences: Burkavd 1 and Xantavd. **Legavidins** and **bradavidins 3** proved interesting for they presented multiple changes similar to the features of the **dimeric avidins**. **Metavidins** are probably the most different of all the putative clades, when the biotin-binding and conserved positions are considered. The **β6 avidins** and **extended avidins** showed several mutual changes and could show similar properties upon experimental inspection.

There were also several outlier sequences within the sub groups that could be of special interest, as well as two sequences that did not fit any sub group. The sequence ENDO ELYS0 is one rather unique outlier sequence in the clade **burkavidins 2**. The substitutions of Ser88Ala, Tyr43Phe, Leu56Tyr, Leu110Trp, and Trp120Phe, are the most notable changes within this sequence. The spatial effects of the substitution of Ser88Ala and Leu110Trp are illustrated in the Figure 24 panels B1–B2 and D1 and D3, respectively. The three outlier sequences in the **burkavidins 2** clade should be considered for further research as well. These sequences do not show homology with any of the usually conserved AA positions in the β8-strand. Furthermore, the Asn23 and Ser27 are both replaced by Gly residues. Two outlier sequences in the **legavidins** clade, ENDO ELYS1 and ENDO MONT, could be interesting research targets as well. Aside from the Trp75Asn substitution (illustrated in Fig. 24 panels A1–A2), these sequences contain generally as much conservation in the conserved positions of avidin family as the clades with verified avidin sequences.

The two sequences left outside of the sub groups, RHOD SP01 and AMIN CIRC, were mostly avidin-like. The firstone resembled the **extended avidins** sub group in sequence and also contained a similar C-terminal extension. However, the latter did not resemble any of the sub groups closely. Most of the biotin-binding positions were conserved, except for Asn23Asp, Tyr45His, Ser88Phe, and Trp92Thr. Thus AMIN CIRC sequence could be interesting target for experimental verification.

6.2.4 Larger scale changes in sequences

Two of the putative avidin branches contained large sequence aberrations compared to the verified avidins. The **β6 avidins** contained rather unusual sequence in the β6-strand and two of the four sequences in this group also seemed to contain an N-terminal extension. The extended avidins on the other hand all contained a C-terminal extension while 2 sequences, PSEU VERO

and ALIA MARI, were extended on the N-terminus as well. The N-terminal extensions did not align together and may not be part of the mature sequences.

The unusual β6-strand sequence in the **β6 avidins** contained substitutions of the Ser88Asp and Thr90Ser. Furthermore, there was an insertion between the Thr90 and the Trp92. Both residues between them, a Trp and a His, were well conserved. The effects of this unusual β6-strand sequence should be further investigated, either by verifying the clades' sequences experimentally or constructing a recombinant avidin with a corresponding β6-strand. Interestingly, these changes are also present in the **extended avidins** clade. Since these clades shared few similar one AA changes, as well, there could be some shared history between the groups' sequences.

Unlike the **β6 avidins**, the **extended avidins** contained further large scale sequence aberrations in the two last β-strands. In the β7-strand, **the extended avidins** featured a substitution of the Thr106 to a variety of hydrophobic residues and even more notably a non-conservative change of Trp108His or Asn. The exact effects of these changes in the context of the whole protein are impossible to decipher without experimental procedures, however, the Trp108 substitutions have been measured to lower the Streptavd's affinity to biotin (Laitinen et al. 2006). Additionally, the Trp108His substitution was illustrated to assess the special effects in the β-barrel (illustrated in Fig. 24 panels C1–C2).

The changes in the β8-strand were even greater. The sequences contained likely homologues for the three key AA positions in this strand: Gly126, Asp128, and Phe130 or a conservative substitution of Phe130Tyr. However, the β-strand in itself seemed elongated in sequence. There were two additional AAs present between the Gly126 and Asp128 pair, but also between the Asp128 and Phe130 pair. As with the **β6 avidins**, sequence from the **extended avidins** clade should be verified or excluded as an avidin. To assess the effects of the unusual β7- and β8-strand sequences a recombinant AVD or Streptavd could be constructed.

Aside from the actual avidin-like segment, the **extended avidins** contained a long C-terminal tail. The tail sequences aligned against each other and weakly against the avidin sequences, as well. However, the quality of the alignment against avidins was poor and could be considered a coincidence. The BLAST queries with the tail sequences against the NCBI databases did not yield any protein results either and thus the origin of these sequences remains elusive. Although the **extended avidins** sequences could be explained by a fusion event with another protein, the fusion partner cannot be identified. Another possibility is a duplication event with the N-terminal avidin copy having undergone more rapid mutation rate. This could, hypothetically, lead to an avidin with two β-barrel structures in one subunit. Even if the other

barrel would not functionally bind biotin, the pseudodimeric structure with one binding site could have biological function for the expressing organism.

Yet, the possibility that these sequences are not true functional avidin family members needs to be considered. Similar to these **extended avidins**, the fibropellin protein from the sea urchin was considered as a possibly functional avidin, when first discovered (Bisgrove et al. 1995). While it is significantly longer than the avidin proteins, fibropellin contains a sequence that closely resembles that of the avidin family. However, this avidin-like domain serves solely as a structural element. As fibropellins, this group of new putative avidins might represent a whole new kind of protein that simply utilizes the tight and effective avidin fold as a building block. Even the possibility of these sequences representing a non-functional pseudogene needs to be taken into account, until they can be experimentally produced or a functional promoter is identified.

6.2.5 Avidin gene evolution in bacteria

The avidin family is likely evolutionarily old as it is present across the oviparous and recently non-chordate species, fungi and prokaryotes. The idea is further supported by the low sequence identities and similarities between the verified avidins. The results in this thesis support the idea further showing that the avidins are possibly even more widely spread across prokaryotes and varied in sequence than previously thought.

Another explanation for the interesting spread of avidin gene would be the HGT. The recent evidence shows it can occur also between bacteria and eukaryotes, especially in an environment where the involved species have a close relationship. Perhaps a laid egg could serve as an opportunity for such event as it is vulnerable to the environment and the transferred gene would directly end up to its descendants. HGT could also explain why some of the branches in the putative avidins tree contained sequences from very heterogenic group of origin species. However, it is impossible to say for certain, as avidins generally have such low identities that the speciation in distant past might be just as possible explanation. Perhaps the only group with very distinctly unrelated origin species and highly conserved sequences was **β6 avidins**.

It is also interesting to note the bacterial species with avidin are mostly from very heterogenic environments. In line with this observation, most of the origin species with fully sequenced genome available contained either more than one chromosome, plasmids, or both (data not shown). Furthermore, avidin was found from the plasmid or secondary chromosome as well as from the primary chromosome. To name some example, a putative avidin sequence,

LEGI PNEU, was present in the primary chromosome of *Legionella pneumophila* (GenBank, Assembly: GCA_000048645) despite the species containing a plasmid and the sequences BURK PSEU0 and BURK PSEU1 originated both from the primary chromosome of the *Burkholderia pseudomallei* (GenBank, Assembly: GCA_000959225) although the species contains also a secondary one. In some of the species avidins are present as multiple copies. Although avidin is still present primarily as an adaptive gene, it sometimes has been essential enough to have been incorporated into the primary replicon of the bacteria, sometimes even in more than one copy. Yet the actual biological function of avidin is still a partially open question.

6.3 Enrichment analysis

The enrichment analysis was performed to assess the avidin gene association with other genes. The goal was to investigate the possible synteny or arrangement into operons. However, the results were rather inconclusive. The GO terms for each gene were fetched and the Fischer exact test was performed to assess whether the term related genes were present more often at avidin vicinity than not.

As a slight disappointment, there was only three terms directly related to defence pathways, GO:0051607 (defense response to virus), GO:0005976 (virion) and GO:0006950 (response to stress). Two of these were enriched with p-value less than 0.05 and last with a p-value of 0.17. Instead, the results showed 14 DNA processing and mobile elements related GO-terms enriched with p-value less than 0.1 and altogether the list contained 19 GO-terms related to these functions. Also basic metabolic pathways were highly represented within the enriched GO-terms; overall 11 terms related to these functions were found near the avidin gene and 7 of these had p-values below 0.1.

The DNA processing and mobile elements related terms could be explained by the association to plasmid mobilization genes as some of the avidin genes originated from plasmids. Also the core metabolic pathways could be enriched by random effect when the avidin gene has been present in the primary chromosome. In bacteria, it is common for the primary chromosome to contain all the house-keeping genes. To assess, whether avidin gene has true co-occurrence with the genes in either of these categories, the avidin genes residing in plasmids and in primary chromosomes should be evaluated independently.

The strict definition of the avidin gene's vicinity could be reconsidered as well. In this study, a threshold of 500 bp up and downstream of the avidin gene was used to select the genes near the avidin gene. As the average bacterial gene length is 900 bp and there is only little space between coding genes in bacterial genomes, this limitation results in the first up and

downstream genes from avidin to be considered. More significant and representative results may be obtained if longer span, such as 2500 bp or 5000 bp is used.

6.4 Tamavidin origin

Against the expectations, Tamavd 1 and Tamavd 2 clustered together with streptavidins instead of other eukaryotic avidins. This is consistent with an earlier thesis work, where the fungal tamavidins clustered together with the streptavidins in both phylogenetic and principal component analyses. As the tamavidins were directly isolated form a market variety *P. cornucopiae*, it is possible the mushroom would have contained symbiotic bacterial species or contamination of soil bacteria.

To further address this idea, avidin genes were sought with BLAST from the fungal genomes. However, no other fungal avidins could be found. Yet, as only a few fungal full genomes were available, the result is not conclusive. No full sequence hits were found from protein or non-genomic databases either. Inspection of codon use in tamavidin did not reveal any conclusive results for neither bacterial nor eukaryotic origin.

HGT is also a possible explanation for the unusual similarity between streptavidins and tamavidins, since HGT has been shown to occur relatively often from bacteria to fungi in rhizosphere. However, verification of this suggestion is difficult. Furthermore, a newly discovered fungal avidin pair, lentiavidin 1 and lentiavidin 2, from *L. edodes* changes the picture (Takakura et al. 2016). This avidin was reported and verified after the phylogenetic analysis was concluded and hence it is not included in the results of this work. If lentiavidins show close relation to tamavidins, the HGT hypothesis becomes less likely, as the HGT would have had to happen from bacteria to both *P. cornucopiae* and *L. edodes*, or alternatively first from bacteria to one fungal species and later between the fungi. Further study to place the lentiavidins in this phylogenetic landscape is required.

7 Conclusions

Altogether 946 protein and 213 nucleotide sequences of putative avidins were identified and collected to an avidin database, DATAvidin. From these, a set of 112 protein sequences these were chosen to be used together with the 14 verified avidin sequences for further analyses. A structural MSA was constructed of the verified avidin sequences. This MSA was used as a profile for constructing a more comprehensive MSA including the 112 putative avidins. For both MSAs a phylogenetic ML tree was built. Finally, ten origin genomes among the 112 sequence set were chosen for the enrichment analysis. The genomes were tested for the avidin genes' association with genes related to specific GO-terms with Fischer's exact test.

A wide set of new putative avidins were identified. Some of these sequences were from species that have not previously been reported to contain avidin genes. The new database, DATAvidin, is hoped to serve as a hub of avidin sequences and help in choosing and comparing potential research targets. The MSAs and the complementary phylogenetic trees illustrated even greater than expected variety among avidin sequences and origin species. These results will be further published as a scientific article. The large scale MSA further revealed that the avidins as a family barely contain AA positions that are not prone to substitutions. Instead, the general similarity of the sequences might be the key to finding a signature features of avidin family.

The genomic context of avidin seemed to tie to mobile elements or house-keeping genes, however, upon further inspection the connection could be coincidental. The correlation could be caused by some of the avidin genes residing in plasmids and others in the main chromosome. Thus no clue about the biological function of avidin could be found aside the origin species habitats. To clarify this, a further enrichment should be performed with separated study sets for the plasmid derived and chromosomal avidins. The species generally preferred highly variable environments with close interactions with higher organisms. This fits the idea of avidin playing a role as either mutualistically beneficial or competitive protein between the host and the bacteria.

Altogether, more detailed research is required to verify the new putative avidins. All sequences from the putative avidin clades, **burkavidins 1**, **metavidins**, **legavidins**, **bradavidins 3**, **β6 avidins**, and **extended avidins**, are a fruitful ground for further investigation. Individual outlier sequences within and outside of the clades were especially interesting targets, as well. Some of the previously unreported AA changes were discovered in these sequences and included, but are not restricted to, Trp75Asn, Ser88Ala, Trp108His, Leu110Met, and Leu110Trp.

References

AHLROTH, M.K., GRAPPUTO, A., LAITINEN, O.H. and KULOMAA, M.S. (2001). Sequence features and evolutionary mechanisms in the chicken avidin gene family. *Biochemical and Biophysical Research Communications,* 285(3), pp. 734-741.

AHLROTH, M.K., KOLA, E.H., KULOMAA, M.S., EWALD, D., MASABANDA, J., SAZANOV, A. and FRIES, R. (2000). Characterization and chromosomal localization of the chicken avidin gene family. *Animal Genetics,* 31(6), pp. 367-375.

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. and LIPMAN, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology,* 215(3), pp. 403-410.

ARMOUGOM, F., MORETTI, S., POIROT, O., AUDIC, S., DUMAS, P., SCHAELI, B., KEDUAS, V. and NOTREDAME, C. (2006). Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic acids research,* 34, pp. W608.

ASHBURNER, M., BALL, C.A., BLAKE, J.A., BOTSTEIN, D., BUTLER, H., CHERRY, J.M., DAVIS, A.P., DOLINSKI, K., DWIGHT, S.S., EPPIG, J.T., HARRIS, M.A., HILL, D.P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J.C., RICHARDSON, J.E., RINGWALD, M., RUBIN, G.M. and SHERLOCK, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics,* 25(1), pp. 25-29.

AVRAHAM, O., MEIR, A., FISH, A., BAYER, E.A. and LIVNAH, O. (2015). Hoefavidin: A dimeric bacterial avidin with a C-terminal binding tail. *Journal of structural biology,* 191(2), pp. 139-148.

BAYER, E.A., KULIK, T., ADAR, R. and WILCHEK, M. (1995). Close similarity among streptavidin-like, biotin-binding proteins from Streptomyces. *Biochimica et biophysica,* 1263(1), pp. 60-66.

BENSON, D.A., KARSCH-MIZRACHI, I., LIPMAN, D.J., OSTELL, J. and WHEELER, D.L. (2005). GenBank. *Nucleic Acids Research,* 33, pp. D38.

BERMAN, H.M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T.N., WEISSIG, H., SHINDYALOV, I.N. and BOURNE, P.E. (2000). The Protein Data Bank. *Nucleic Acids Research,* 28(1), pp. 235-242.

BISGROVE, B.W., ANDREWS, M.E. and RAFF, R.A. (1995). Evolution of the fibropellin gene family and patterns of fibropellin gene expression in sea urchin phylogeny. *Journal of Molecular Evolution,* 41(1), pp. 34-45.

BLEULER-MARTINEZ, S., SCHMIEDER, S., AEBI, M. and KÜNZLER, M. (2012). Biotin-binding proteins in the defense of mushrooms against predators and parasites. *Applied and Environmental Microbiology,* 78(23), pp. 8485-8487.

BORATYN, G.M., SCHÄFFER, A.A., AGARWALA, R., ALTSCHUL, S.F., LIPMAN, D.J. and MADDEN, T.L. (2012). Domain enhanced lookup time accelerated BLAST. *Biology Direct,* 7(1), pp. 12.

BOTTE, V. and GRANATA, G. (1977). Induction of avidin synthesis by RNA obtained from lizard oviducts. *The Journal of Endocrinology,* 73(3), pp. 535-536.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. and MADDEN, T.L. (2009). BLAST+: architecture and applications. *BMC bioinformatics,* 10, pp. 421.

CASJENS, S. (1998). The diverse and dynamic structure of bacterial genomes. *Annual Review of Genetics,* 32, pp. 339-377.

CHAIET, L. and WOLF, F.J. (1964). The properties of streptavidin, a biotin-binding protein produced by Streptomycetes. *Archives of Biochemistry and Biophysics,* 106, pp. 1-5.

CHANG, G.S., HONG, Y., KO, K.D., BHARDWAJ, G., HOLMES, E.C., PATTERSON, R.L. and VAN ROSSUM, D.B. (2008). Phylogenetic Profiles Reveal Evolutionary Relationships within the "Twilight Zone" of Sequence Similarity. *Proceedings of the National Academy of Sciences of the United States of America,* 105(36), pp. 13474-13479.

DARMON, E. and LEACH, D.R.F. (2014). Bacterial Genome Instability. *Microbiology and Molecular Biology Reviews,* **7**8(1), pp. 1-39.

DELAMUTA, J.R., RIBEIRO, R.A., ORMENO-ORRILLO, E., MELO, I.S., MARTINEZ-ROMERO, E. and HUNGRIA, M. (2013). Polyphasic evidence supporting the reclassification of Bradyrhizobium japonicum group Ia strains as Bradyrhizobium diazoefficiens sp. nov. *International Journal of Systematic and Evolutionary Microbiology,* 63(9), pp. 3342-3351.

DUNNING HOTOPP, J.C. (2011). Horizontal gene transfer between bacteria and animals. *Trends in genetics,* 27(4), pp. 157-163.

EAKIN, R.E., MCKINLEY, W.A. and WILLIAMS, R.J. (1940). Egg-White Injury in Chicks and Its Relationship to a Deficiency of Vitamin H (Biotin). *Science,* 92, pp. 224-225.

EDGAR, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32(5), pp. 1792-1797.

EFRON, B., HALLORAN, E. and HOLMES, S. (1996). Bootstrap Confidence Levels for Phylogenetic Trees. *Proceedings of the National Academy of Sciences of the United States of America,* 93(14), pp. 7085-7090.

EISENBERG-DOMOVICH, Y., HYTÖNEN, V.P., WILCHEK, M., BAYER, E.A., KULOMAA, M.S. and LIVNAH, O. (2005). High-resolution crystal structure of an avidin-related protein: insight into high-affinity biotin binding and protein stability. *Acta crystallographica,* 61(5), pp. 528-538.

FELSENSTEIN, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution,* 39, pp. 783-791.

FINN, R.D., ATTWOOD, T.K., BABBITT, P.C., BATEMAN, A., BORK, P., BRIDGE, A.J., CHANG, H., DOSZTÁNYI, Z., EL-GEBALI, S., FRASER, M., GOUGH, J., HAFT, D., HOLLIDAY, G.L., HUANG, H., HUANG, X., LETUNIC, I., LOPEZ, R., LU, S., MARCHLER-BAUER, A., MI, H., MISTRY, J., NATALE, D.A., NECCI, M., NUKA, G., ORENGO, C.A., PARK, Y., PESSEAT, S., PIOVESAN, D., POTTER, S.C., RAWLINGS, N.D., REDASCHI, N., RICHARDSON, L., RIVOIRE, C., SANGRADOR-VEGAS, A., SIGRIST, C., SILLITOE, I., SMITHERS, B., SQUIZZATO, S., SUTTON, G., THANKI, N., THOMAS, P.D., TOSATTO, S.C.E., WU, C.H., XENARIOS, I., YEH, L., YOUNG, S. and MITCHELL, A.L. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research,* 45(1), pp. D199.

FINN, R.D., COGGILL, P., EBERHARDT, R.Y., EDDY, S.R., MISTRY, J., MITCHELL, A.L., POTTER, S.C., PUNTA, M., QURESHI, M., SANGRADOR-VEGAS, A., SALAZAR, G.A., TATE, J. and BATEMAN, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research,* 44(1), pp. D285.

FLOWER, D.R. (1996). The lipocalin protein family: structure and function. *Biochemical Journal,* 318(1), pp. 1-14.

FLOWER, D.R. (1993). Structural relationship of streptavidin to the calycin protein superfamily. *FEBS letters,* 333(1-2), pp. 99-102.

FLOWER, D.R., NORTH, A.C. and SANSOM, C.E. (2000). The lipocalin protein family: structural and sequence overview. *Biochimica et biophysica acta,* 1482(1-2), pp. 9-24.

GAO, C., REN, X., MASON, A.S., LIU, H., XIAO, M., LI, J. and FU, D. (2014). Horizontal gene transfer in plants. *Functional & Integrative Genomics,* 14(1), pp. 23-29.

GREEN, N.M. (1990). Avidin and streptavidin. *Methods in Enzymology,* 184, pp. 51-67.

GREEN, N.M. (1975). Avidin. *Advances in Protein Chemistry,* 29, pp. 85-133.

GREEN, N.M. (1963). Avidin. 1. The use of (14-C)biotin for kinetic studies and for assay. *The Biochemical Journal,* 89, pp. 585-591.

GUO, X., XIN, J., WANG, P., DU, X., JI, G., GAO, Z. and ZHANG, S. (2017). Functional characterization of avidins in amphioxus Branchiostoma japonicum: Evidence for a dual role in biotin-binding and immune response. *Developmental and Comparative Immunology,* 70, pp. 106-118.

HAFT, D.H., SELENGUT, J.D. and WHITE, O. (2003). The TIGRFAMs database of protein families. *Nucleic acids research,* 31(1), pp. 371-373.

HELPPOLAINEN, S.H., MÄÄTTÄ, J.A., HALLING, K.K., SLOTTE, J.P., HYTÖNEN, V.P., JÄNIS, J., VAINIOTALO, P., KULOMAA, M.S. and NORDLUND, H.R. (2008). Bradavidin II from Bradyrhizobium japonicum: a new avidin-like biotin-binding protein. *Biochimica et biophysica acta,* 1784(7-8), pp. 1002-1010.

HELPPOLAINEN, S.H., NURMINEN, K.P., MÄÄTTÄ, J.A.E., HALLING, K.K., SLOTTE, J.P., HUHTALA, T., LIIMATAINEN, T., YLÄ-HERTTUALA, S., AIRENNE, K.J., NÄRVÄNEN, A., JÄNIS, J., VAINIOTALO, P., VALJAKKA, J., KULOMAA, M.S. and NORDLUND, H.R. (2007). Rhizavidin from Rhizobium etli: the first natural dimer in the avidin protein family. *The Biochemical journal,* 405(3), pp. 397-405.

HENDRICKSON, W.A., PAHLER, A., SMITH, J.L., SATOW, Y., MERRITT, E.A. and PHIZACKERLEY, R.P. (1989). Crystal Structure of Core Streptavidin Determined from Multiwavelength Anomalous Diffraction of Synchrotron Radiation. *Proceedings of the National Academy of Sciences of the United States of America,* 86(7), pp. 2190-2194.

HERTZ, R. and SEBRELL, W.H. (1942). Occurrence of avidin in the oviduct and secretions of the genital tract of several species. *Science,* 96, pp. 257.

HOOD, E.E., WITCHER, D.R., MADDOCK, S., MEYER, T., BASZCZYNSKI, C., BAILEY, M., FLYNN, P., REGISTER, J., MARSHALL, L., BOND, D., KULISEK, E., KUSNADI, A., EVANGELISTA, R., NIKOLOV, Z., WOOGE, C., MEHIGH, R.J., HERNAN, R., KAPPEL, W.K., RITLAND, D., LI, C.P. and HOWARD, J.A. (1997). Commercial production of avidin from transgenic maize: characterization of transformant, production, processing, extraction and purification. *Molecular Breeding,* 3(4), pp. 291-306.

HYRE, D.E., AMON, L.M., PENZOTTI, J.E., LE TRONG, I., STENKAMP, R.E., LYBRAND, T.P. and STAYTON, P.S. (2002). Early mechanistic events in biotin dissociation from streptavidin. *Nature Structural Biology,* 9(8), pp. 582-585.

HYTÖNEN, V.P., MÄÄTTÄ, J.A., KIDRON, H., HALLING, K.K., HÖRHÄ, J., KULOMAA, T., NYHOLM, T.K., JOHNSON, M.S., SALMINEN, T.A., KULOMAA, M.S. and AIRENNE, T.T. (2005). Avidin related protein 2 shows unique structural and functional features among the avidin protein family. *BMC biotechnology,* 5, pp. 28.

HYTÖNEN, V.P., MÄÄTTÄ, J.A., NISKANEN, E.A., HUUSKONEN, J., HELTTUNEN, K.J., HALLING, K.K., NORDLUND, H.R., RISSANEN, K., JOHNSON, M.S., SALMINEN, T.A., KULOMAA, M.S., LAITINEN, O.H. and AIRENNE, T.T. (2007). Structure and characterization of a novel chicken biotin-binding protein A (BBP-A). *BMC structural biology,* 7, pp. 8.

HYTÖNEN, V.P., LAITINEN, O.H., GRAPPUTO, A., KETTUNEN, A., SAVOLAINEN, J., KALKKINEN, N., MARTTILA, A.T., NORDLUND, H.R., NYHOLM, T.K.M., PAGANELLI, G. and KULOMAA, M.S. (2003). Characterization of poultry egg-white avidins and their potential as a tool in pretargeting cancer treatment. *The Biochemical Journal,* 372(1), pp. 219-225.

JONES, D.T., TAYLOR, W.R. and THORNTON, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences,* 8(3), pp. 275-282.

KANZ, C., ALDEBERT, P., ALTHORPE, N., BAKER, W., BALDWIN, A., BATES, K., BROWNE, P., VAN DEN BROEK, A., CASTRO, M., COCHRANE, G., DUGGAN, K., EBERHARDT, R., FARUQUE, N., GAMBLE, J., DIEZ, F.G., HARTE, N., KULIKOVA, T., LIN, Q., LOMBARD, V., LOPEZ, R., MANCUSO, R., MCHALE, M., NARDONE, F. and SILVENTOINEN (2005). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research,* 33, pp. 33.

KEELING, P.J. and PALMER, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews. Genetics,* 9(8), pp. 605-618.

KEINÄNEN, R.A., WALLÉN, M.J., KRISTO, P.A., LAUKKANEN, M.O., TOIMELA, T.A., HELENIUS, M.A. and KULOMAA, M.S. (1994). Molecular cloning and nucleotide sequence of chicken avidin-related genes 1-5. *European Journal of Biochemistry,* 220(2), pp. 615-621.

KIM, S., THIESSEN, P.A., BOLTON, E.E., CHEN, J., FU, G., GINDULYTE, A., HAN, L., HE, J., HE, S., SHOEMAKER, B.A., WANG, J., YU, B., ZHANG, J. and BRYANT, S.H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Research,* 44(1), pp. 1202.

KOONIN, E.V., MAKAROVA, K.S. and ARAVIND, L. (2001). Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annual Review of Microbiology*, 55, pp. 709-742.

KORPELA, J.K., ELO, H.A. and TUOHIMAA, P.J. (1981). Avidin induction by estrogen and progesterone in the immature oviduct of chicken, Japanese Quail, duck, and gull. *General and Comparative Endocrinology,* 44(2), pp. 230-232.

KORPELA, J., KULOMAA, M., TUOHIMAA, P. and VAHERI, A. (1982). Induction of avidin in chickens infected with the acute leukemia virus OK 10. *International Journal of Cancer,* 30(4), pp. 461-464.

KUNNAS, T.A., WALLÉN, M.J. and KULOMAA, M.S. (1993). Induction of chicken avidin and related mRNAs after bacterial infection. *Biochimica et biophysica acta,* 1216(3), pp. 441-445.

KURZBAN, G.P., BAYER, E.A., WILCHEK, M. and HOROWITZ, P.M. (1991). The quaternary structure of streptavidin in urea. *The Journal of Biological Chemistry,* 266(22), pp. 14470-14477.

LAITINEN, O.H., HYTÖNEN, V.P., AHLROTH, M.K., PENTIKÄINEN, O.T., GALLAGHER, C., NORDLUND, H.R., OVOD, V., MARTTILA, A.T., PORKKA, E., HEINO, S., JOHNSON, M.S., AIRENNE, K.J. and KULOMAA, M.S. (2002). Chicken avidin-related proteins show altered biotin-binding and physico-chemical properties as compared with avidin. *The Biochemical Journal,* 363(3), pp. 609-617.

LAITINEN, O.H., HYTÖNEN, V.P., NORDLUND, H.R. and KULOMAA, M.S. (2006). Genetically engineered avidins and streptavidins. *Cellular and molecular life sciences,* 63(24), pp. 2992-3017.

LARSSON, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics,* 30(22), pp. 3276-3278.

LEPPINIEMI, J., GRÖNROOS, T., MÄÄTTÄ, J.A., JOHNSON, M.S., KULOMAA, M.S., HYTÖNEN, V.P. and AIRENNE, T.T. (2012). Structure of bradavidin-C-terminal residues act as intrinsic ligands. *PloS one,* 7(5), pp. e35962.

LEPPINIEMI, J., MEIR, A., KÄHKÖNEN, N., KUKKURAINEN, S., MÄÄTTÄ, J.A., OJANEN, M., JÄNIS, J., KULOMAA, M.S., LIVNAH, O. and HYTÖNEN, V.P. (2013). The highly dynamic oligomeric structure of bradavidin II is unique among avidin proteins. *Protein science,* 22(7), pp. 980-994.

LEPPINIEMI, J., MÄÄTTÄ, J.A.E., HAMMAREN, H., SOIKKELI, M., LAITAOJA, M., JÄNIS, J., KULOMAA, M.S. and HYTÖNEN, V.P. (2011). Bifunctional avidin with covalently modifiable ligand binding site. *PloS one,* 6(1), pp. e16576.

LIÒ, P. (2002). Investigating the Relationship Between Genome Structure, Composition, and Ecology in Prokaryotes. *Molecular Biology and Evolution,* 19(6), pp. 789-800.

LIVNAH, O., BAYER, E.A., WILCHEK, M. and SUSSMAN, J.L. (1993). Three-dimensional structures of avidin and the avidin-biotin complex. *Proceedings of the National Academy of Sciences of the United States of America,* 90(11), pp. 5076-5080.

MÄÄTTÄ, J.A., HELPPOLAINEN, S.H., HYTÖNEN, V.P., JOHNSON, M.S., KULOMAA, M.S., AIRENNE, T.T. and NORDLUND, H.R. (2009). Structural and functional characteristics of xenavidin, the first frog avidin from Xenopus tropicalis. *BMC structural biology,* 9, pp. 63.

MARTIN, H., BURGESS, E.P.J., MASARIK, M., KRAMER, K.J., BEKLOVA, M., ADAM, V. and KIZEK, R. (2010). Avidin and Plant Biotechnology to Control Pests. *Genetic Engineering, Biofertilisation, Soil Quality and Organic Farming*, pp. 1-21.

MASHIMA, J., KODAMA, Y., FUJISAWA, T., KATAYAMA, T., OKUDA, Y., KAMINUMA, E., OGASAWARA, O., OKUBO, K., NAKAMURA, Y. and TAKAGI, T.(2017). DNA Data Bank of Japan. *Nucleic Acids Research,* 45(1), pp. D31.

MEIR, A., BAYER, E.A. and LIVNAH, O. (2012). Structural adaptation of a thermostable biotin-binding protein in a psychrophilic environment. *The Journal of biological chemistry,* 287(22), pp. 17951-17962.

MEIR, A., HELPPOLAINEN, S.H., PODOLY, E., NORDLUND, H.R., HYTÖNEN, V.P., MÄÄTTÄ, J.A., WILCHEK, M., BAYER, E.A., KULOMAA, M.S. and LIVNAH, O. (2009). Crystal structure of rhizavidin: insights into the enigmatic high-affinity interaction of an innate biotin-binding protein dimer. *Journal of Molecular Biology,* 386(2), pp. 379-390.

NISKANEN, E.A., HYTÖNEN, V.P., GRAPPUTO, A., NORDLUND, H.R., KULOMAA, M.S. and LAITINEN, O.H. (2005). Chicken genome analysis reveals novel genes encoding biotin-binding proteins related to avidin family. *BMC genomics,* 6, pp. 41.

NORDLUND, H.R., HYTÖNEN, V.P., LAITINEN, O.H. and KULOMAA, M.S. (2005). Novel avidin-like protein from a root nodule symbiotic bacterium, Bradyrhizobium japonicum. *The Journal of biological chemistry,* 280(14), pp. 13250-13255.

NOTREDAME, C., HIGGINS, D.G. and HERINGA, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology,* 302(1), pp. 205-217.

O'LEARY, N.A., WRIGHT, M.W., BRISTER, J.R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., FARRELL, C.M., GOLDFARB, T., GUPTA, T., HAFT, D., HATCHER, E., HLAVINA, W., JOARDAR, V.S., KODALI, V.K., LI, W., MAGLOTT, D., MASTERSON, P., MCGARVEY, K.M., MURPHY, M.R., O'NEILL, K., PUJAR, S., RANGWALA, S.H., RAUSCH, D., RIDDICK, L.D., SCHOCH, C., SHKEDA, A., STORZ, S.S., SUN, H., THIBAUD-NISSEN, F., TOLSTOY, I., TULLY, R.E., VATSAN, A.R., WALLIN, C., WEBB, D., WU, W., LANDRUM, M.J., KIMCHI, A., TATUSOVA, T., DICUCCIO, M., KITTS, P., MURPHY, T.D. and PRUITT, K.D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research,* 44(1), pp. 733.

PÁL, C., PAPP, B. and LERCHER, M.J. (2006). An integrated view of protein evolution. *Nature Reviews. Genetics,* 7(5), pp. 337-348.

PEARSON, W.R.(2013). Selecting the Right Similarity-Scoring Matrix. *Current protocols in bioinformatics,* 43, pp. 3.5.1.

PONTING, C.P. and RUSSELL, R.R.(2002). The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure,* 31, pp. 45-71.

PRZYBYLSKI, D. and ROST, B.(2008). Powerful fusion: PSI-BLAST and consensus sequences. *Bioinformatics,* 24(18), pp. 1987-1993.

QIU, H., CAI, G., LUO, J., BHATTACHARYA, D. and ZHANG, N. (2016). Extensive horizontal gene transfers between plant pathogenic fungi. *BMC biology,* 14, pp. 41.

RICHARDSON, A.O. and PALMER, J.D. (2007). Horizontal gene transfer in plants. *Journal of Experimental Botany,* 58(1), pp. 1-9.

ROCHA, E.P.C. (2008). The organization of the bacterial genome. *Annual Review of Genetics,* 42, pp. 211-233.

ROCHA, E.P.C. (2004). The replication-related organization of bacterial genomes. *Microbiology,* 150(6), pp. 1609-1627.

ROSANO, C., AROSIO, P. and BOLOGNESI, M. (1999). The X-ray three-dimensional structure of avidin. *Biomolecular engineering,* 16(1-4), pp. 5-12.

SAITOU, N. and NEI, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* 4(4), pp. 406-425.

SARDO, A., WOHLSCHLAGER, T., LO, C., ZOLLER, H., WARD, T.R. and CREUS, M. (2011). Burkavidin: a novel secreted biotin-binding protein from the human pathogen Burkholderia pseudomallei. *Protein expression and purification,* 77(2), pp. 131-139.

SINKKONEN, A., LAITINEN, O.H., LEPPINIEMI, J., VAURAMO, S., HYTÖNEN, V.P. and SETÄLÄ, H. (2014). Positive association between biotin and the abundance of root-feeding nematodes. *Soil Biology and Biochemistry,* 73, pp. 93-95.

SMILLIE, C., GARCILLÁN-BARCIA, M.P., FRANCIA, M.V., ROCHA, E.P.C. and DE LA CRUZ, F. (2010). Mobility of Plasmids. *Microbiology and Molecular Biology Reviews,* 74(3), pp. 434-452.

STREIT, W.R. and ENTCHEVA, P. (2003). Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. *Applied Microbiology and Biotechnology,* 61(1), pp. 21-31.

SULLIVAN, J. (2005). Maximum-likelihood methods for phylogeny estimation. *Methods in Enzymology,* 395, pp. 757-779.

TAKAKURA, Y., OKA, N., SUZUKI, J., TSUKAMOTO, H. and ISHIDA, Y. (2012). Intercellular production of tamavidin 1, a biotin-binding protein from Tamogitake mushroom, confers resistance to the blast fungus Magnaporthe oryzae in transgenic rice. *Molecular biotechnology,* 51(1), pp. 9-17.

TAKAKURA, Y., SOFUKU, K., TSUNASHIMA, M. and KUWATA, S. (2016). Lentiavidins: Novel avidin-like proteins with low isoelectric points from shiitake mushroom (Lentinula edodes). *Journal of bioscience and bioengineering,* 121(4), pp. 420-423.

TAKAKURA, Y., TSUNASHIMA, M., SUZUKI, J., USAMI, S., KAKUTA, Y., OKINO, N., ITO, M. and YAMAMOTO, T. (2009). Tamavidins--novel avidin-like biotin-binding proteins from the Tamogitake mushroom. *The FEBS journal,* 276(5), pp. 1383-1397.

TAMURA, K., STECHER, G., PETERSON, D., FILIPSKI, A. and KUMAR, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution,* 30(12), pp. 2725-2729.

TASKINEN, B., ZMURKO, J., OJANEN, M., KUKKURAINEN, S., PARTHIBAN, M., MÄÄTTÄ, J.A., LEPPINIEMI, J., JÄNIS, J., PARIKKA, M., TURPEINEN, H., RÄMET, M., PESU, M., JOHNSON, M.S., KULOMAA, M.S., AIRENNE, T.T. and HYTÖNEN, V.P. (2013). Zebavidin--an avidin-like protein from zebrafish. *PloS one,* 8(10), pp. e77207.

TAUSIG, F. and WOLF, F.J. (1964). Streptavidin--a substance with avidin-like properties produced by microorganisms. *Biochemical and Biophysical Research Communications,* 14, pp. 205-209.

THE UNIPROT CONSORTIUM (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research,* 45(D1), pp. D169.

TIWARI, A. (2015). Phylogenetic Analysis of Avidins and Expression of Novel Avidins from Lactrodectus hesperus and Hoeflea phototrophica, University of Turku.

TRANTER, H.S. and BOURD, R.G. (1982). The antimicrobial defense of avian eggs: Biological perspective and chemical basis. *Journal of Applied Biochemistry,* 45.(4), pp. 295-338.

TUOHIMAA, P., JOENSUU, T., ISOLA, J., KEINÄNEN, R., KUNNAS, T., NIEMELÄ, A., PEKKI, A., WALLÉN, M., YLIKOMI, T. and KULOMAA, M. (1989). Development of progestin-specific response in the chicken oviduct. *The International Journal of Developmental Biology,* 33(1), pp. 125-134.

VENESKOSKI, K. (2009). Strongavidiini – Biotiinia sitova merisiilestä peräisin oleva avidiinin
kaltainen proteiini, University of Jyväskylä.

VOGEL, C., BASHTON, M., KERRISON, N.D., CHOTHIA, C. and TEICHMANN, S.A. (2004). Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology,* 14(2), pp. 208-216.

WALLÉN, M.J., LAUKKANEN, M.O. and KULOMAA, M.S. (1995). Cloning and sequencing of the chicken egg-white avidin-encoding gene and its relationship with the avidin-related genes Avr1-Avr5. *Gene,* 161(2), pp. 205-209.

WATERMAN, M.S., SMITH, T.F. and BEYER, W.A. (1976). Some biological sequence metrics. *Advances in Mathematics,* 20(3), pp. 367-387.

WET, R. and HSU, C. (1970). Order of Laying and Avidin Content of Hens' Eggs. *Poultry Science,* 49(2), pp. 517-518.

XI, Z., BRADLEY, R.K., WURDACK, K.J., WONG, K., SUGUMARAN, M., BOMBLIES, K., REST, J.S. and DAVIS, C.C. (2012). Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC genomics,* 13, pp. 227.

YANAI, I., YU, Y., ZHU, X., CANTOR, C.R. and WENG, Z. (2005). An avidin-like domain that does not bind biotin is adopted for oligomerization by the extracellular mosaic protein fibropellin. *Protein science,* 14(2), pp. 417-423.

YOZA, K., IMAMURA, T., KRAMER, K.J., MORGAN, T.D., NAKAMURA, S., AKIYAMA, K., KAWASAKI, S., TAKAIWA, F. and OHTSUBO, K. (2005). Avidin expressed in transgenic rice confers resistance to the stored-product insect pests Tribolium confusum and Sitotroga cerealella. *Bioscience, Biotechnology, and Biochemistry,* 69(5), pp. 966-971.

Appendices

Appendix A. Bacterial species of origin for the putative and verified avidin sequences.

Table Appendix A. The origin bacterial species of the putative avidin sequences.

Sequence	Source organism	Environment niche	Interspecies relations	Genomic location	Source
ACTI ENZA	*Actinokineospora enzanensis*	soil	-	–	BioSample Accession: SAMN02441015
ACTI SERA	*Actinocatenispora sera*	soil	endosporulation	–	DOI: 10.1099/ijs.0.65270-0
AFIF PFEN	*Afifella pfennigii* (former *Rhodobium pfennigii*)	brackish water, benthic zone	microbial mat participant	–	www.bacterio.net Search: afifella
ALIA MARI	*Aliagarivorans marinus*	marine	–	–	DOI: 10.1099/ijs.0.008235-0
AMIN CIRC	*Aminiphilus circumscriptus*	waste sludge	–	–	DOI: 10.1099/ijs.0.63614-0
avidin	*Gallus gallus*	chicken	–	–	
BRAD DIAZ 0, 1	*Bradyrhizobium diazoefficiens*	soil, nitrogen fixation	legume root nodular symbiont	genomic	DOI: 10.1128/genomeA.01743-16
BRAD ELKA 0, 1	*Bradyrhizobium elkanii*	soil, nitrogen fixation	legume root nodular symbiont	–	PMID: 8285723
BRAD JAPO 0, 1	*Bradyrhizobium diazoefficien*s (former *B. japonicum*)	soil, nitrogen fixation	legume root nodular symbiont	genomic	PMID: 12597279 DOI: 10.1099/ijs.0.049130-0
BRAD JAPO 2	*Bradyrhizobium japonicum*	soil, nitrogen fixation	legume root nodular symbiont	genomic	microbewiki.kenyon.edu Search: Bradyrhizobium japonicum
BRAD SP01	*Bradyrhizobium sp. WSM1253*	soil, nitrogen fixation	legume root nodular symbiont	–	DOI: 10.1186/s40793-015-0115-9
BRAD SP02 0, 1	*Bradyrhizobium sp. WSM3983*	soil, nitrogen fixation	legume root nodular symbiont	–	BioSample Accession: SAMN02440715
BRAD SP03	*Bradyrhizobium sp. STM*	soil, nitrogen fixation	legume root nodular symbiont	–	BioProject Accession: PRJNA162993
BRAD YUAN 0,1	*Bradyrhizobium yuanmingense*	soil, nitrogen fixation	legume root nodular symbiont	–	DOI: 10.1111/j.1574-6968.2008.01169.x
Bradavidin I, Bradavidin II	*Bradyrhizobium diazoefficien*s (former *B. japonicum*)	soil, nitrogen fixation	legume root nodular symbiont	genomic	DOI: 10.1016/j.bbapap.2008.04.010 10.1074/jbc.M414336200
BURK BACT 0, 1	*Burkholderiaceae bacterium*	soil	potential human and animal pathogen	genomic	BioSample Accession: SAMN03340296
BURK CENO	*Burkholderia cenocepacia*	soil, aquatic, aerosol	biofilm, opportunistic human pathogen, antibiotic resistant	genomic	PMID: 16217180
BURK CEPA	*Burkholderia cepacia*	soil, aquatic, aerosol	biofilm, opportunistic human pathogen	genomic	PMID: 16217180
BURK GLUM	*Burkholderia glumae*	soil, aquatic, voluntary dormancy	plant pathogen	genomic	DOI: 10.1094/PDIS-10-13-1024-PDN
BURK MALL	*Burkholderia mallei*	–	human and animal pathogen	genomic	PMID: 18221181

BURK MIMO 0, 1	*Burkholderia mimosarum*	soil, nitrogen fixation	mimosa root nodular symbiont	genomic	DOI: 10.1099/ijs.0.64325-0
BURK NODO	*Burkholderia nodosa*	soil, nitrogen fixation	mimosa root nodular symbiont	genomic	DOI: 10.1099/ijs.0.64873-0
BURK OKLA 0, 1	*Burkholderia oklahomensis*	soil, aquatic, voluntary dormancy	human and animal pathogen	genomic	DOI: 10.1099/ijs.0.63991-0
BURK OXYP	*Burkholderia oxyphila*	soil, acidic	–	genomic	DOI: 10.1099/ijs.0.017368-0
BURK PSEU 0, 1	*Burkholderia pesudomallei*	soil, aquatic, voluntary dormancy	human and animal pathogen	genomic	PMID: 18221181
BURK SP01 0, 1	*Burkholderia sp. TSV202*	soil, aquatic, voluntary dormancy	Possible human and animal pathogen	genomic	BioSample Accession: SAMN02951647
BURK THAI 0, 1	*Burkholderia thailandensis*	soil, aquatic, voluntary dormancy	human and animal pathogen	genomic	DOI: 10.1128/JCM.01585-06
burkavidin	*Burkholderia pesudomallei*	soil, aquatic, voluntary dormancy	human and animal pathogen	genomic	PMID: 18221181 DOI: 10.1016/j.pep.2011.01.003
CATE ACID	*Catenulispora acidiphila*	soil, decomposing	–	–	DOI: 10.1099/ijs.0.63858-0
CUPR PINA	*Cupriavidus pinatubonensis*	volcanic sludge, oxidation, metal resistance	–	genomic	DOI: 10.1099/ijs.0.63922-0
CUPR SP01	*Cupriavidus sp. SK-3*	aquatic sediment, contaminant degradation	–	–	DOI: 10.1128/genomeA.00664-14
ENDO ELYS 0, 1	*Endozoicomonas elysicola*	marine	sea slug endogastric flora, coral and sponge microflora	–	DOI: 10.1016/j.syapm.2006.07.003 10.3354/dao02636
ENDO MONT	*Endozoicomonas montiporae*	marine	coral microflora	–	DOI: 10.1099/ijs.0.014357-0 10.3389/fmicb.2016.00251
ENTE CALV	*Vibrio calviensis* (former *Enterovibrio calviensis*)	marine	–	–	DOI: 10.1099/ijs.0.001990-0 10.1099/00207713-52-2-549
FLEX ROSE	*Flexibacter roseolus*	aquatic	fish pathogen	–	www.bacterio.net Genus: Flexibacter
HAEM MISS	*Haematobacter missouriensis*	soil	possible sepsis causing human pathogen	–	DOI: 10.1128/JCM.01188-06
HELL BALN	*Hellea balneolensis*	marine surface	–	–	DOI: 10.1099/ijs.0.65424-0
HOEF PHOT	*Hoeflea phototrophica*	marine, photosynthetic	–	–	DOI: 10.1099/ijs.0.63958-0 10.4056/sigs.3486982
hoefavidin	*Hoeflea phototrophica*	marine, photosynthetic	–	–	DOI: 10.1016/j.jsb.2015.06.020
HYME SP01	*Hymenobacter sp. AT01-02*	desert soil, UV-resistance, Mn and Fe accumulation	–	–	DOI: 10.1128/genomeA.01701-15
INQU LIMO	*Inquilinus limosus*	aquatic	opportunistic human pathogen	–	DOI: 10.3201/eid1103.041078 10.3201/eid1406.071355
KILO SPON	*Kiloniella spongiae*	marine	sponge microflora	–	DOI: 10.1099/ijs.0.069773-0

KITA GRIS	*Kitasatospora griseola* (former *Streptomyces griseolisporeus*)	soil	secretes exotoxins, antibiotics and antifungal agents	–	DOI: 10.1128/genomeA.00208-15
KORD GWAN	*Kordiimonas gwangyangensis*	marine sediment	–	–	DOI: 10.1099/ijs.0.63684-0
LEGI ANIS	*Legionella anisa*	aquatic, requires Cys	plant pathogen, amoebae intracellular parasite	–	PMID: 3985609 DOI: 10.1371/journal.pone.0159726
LEGI CHER	*Legionella cherrii*	aquatic, nitrogen fixation, requires Cys	human pathogen	–	DOI: 10.1128/AEM.69.1.533-541.2003
LEGI FALL	*Legionella fallonii*	aquatic, aerosol, requires Cys	human pathogen	–	DOI: 10.1099/00207713-51-3-1151
LEGI LANS	*Legionella lansingensis*	aquatic	human pathogen	genomic	PMID: 1401005
LEGI MASS	*Legionella massiliensis*	aquatic	human pathogen	–	DOI: 10.1099/ijs.0.037853-0 10.1128/genomeA.01068-14
LEGI PNEU	*Legionella pneumophila*	aquatic, requires Cys	amoebae intracellular parasite	–	DOI: 10.1078/1438-4221-00139
LEGI TUNI	*Legionella tunisiensis*	aquatic	amoebae intracellular parasite	–	DOI: 10.1099/ijs.0.037853-0
LYSO ANTI	*Lysobacter antibioticus*	soil, aquatic, extremophile	plant microflora, salamander skin microflora, secretes antibiotics and antifungal agents, probable legume root nodular symbiont	–	DOI: 10.1007/s00284-009-9481-0 10.3389/fmicb.2015.01243
MARI MEDI	*Marinomonas mediterranea*	marine, contaminant degradation	probable obligate symbiont to other bacteria	–	DOI: 10.1099/mic.0.26524-0 10.4056/sigs.2545743
MARI POSI	*Marinomonas posidonica*	marine	seaweed microflora	genomic	DOI: 10.4056/sigs.2976373
MARI PURP	*Marichromatium purpuratum*	marine, photosynthetic	–	genomic	www.bacterio.net Genus: Marichromatium
MARI SP01	*Euryhalocaulis caribicus* (former *Maricaulis sp. JL2009*)	–	–	–	DOI: 10.1128/genomeA.00407-13 Taxon ID (NCBI): 1161401
MARI SP02	*Marinomonas sp. MWYL1*	marine	salt marsh grass microflora	–	genome.jgi.doe.gov Search: *Marinomonas* MWYL1
MESO AUST	*Mesorhizobium australicum*	soil, nitrogen fixation	legume root nodular symbiont	genomic	DOI: 10.1099/ijs.0.005728-0
MESO CICE	*Mesorizobium plurifarum* (former *M. ciceri*)	soil, nitrogen fixation	legume root nodular symbiont	genomic	DOI: 10.4056/sigs.4458283 10.1111/j.1574-6941.2009.00776.x
MESO OPPO	*Mesorhizobium opportunistum*	soil, nitrogen fixation	legume root nodular symbiont	–	DOI: 10.4056/sigs.4538264 10.1099/ijs.0.005728-0
MESO PLUR	*Mesorizobium plurifarum* (former *M. ciceri*)	soil, nitrogen fixation	legume root nodular symbiont	genomic	DOI: 10.4056/sigs.4458283 10.1111/j.1574-6941.2009.00776.x

MESO SP04	*Mesorhizobium sp. LSJC280B00*	soil	–	–	BioSample Accession: SAMN02359695
MESO SP06	*Mesorhizobium sp. LSJC255A00*	soil	–	–	BioSample Accession: SAMN02359689
METH EXTO	*Methylobacterium extorquens*	soil, aquatic sediment	opportunistic human pathogen	–	DOI: 10.1371/journal.pone.0013001
METH MESO	*Methylobacterium mesophilicum*	soil	opportunistic human pathogen	plasmid	DOI: 10.1086/313815 PMID: 8469180
METH RADI	*Methylobacterium radiotolerans*	soil, radiation resistance	opportunistic human pathogen	–	microbewiki.kenyon.edu Genus: Metylobacterium DOI: 10.1128/JCM.01241-11
MYCO THER	*Mycobacterium thermoresistibile*	soil, aquatic	human pathogen, found in mucous membranes, urine, gastric fluid	plasmid	DOI: 10.1002/pro.2084 PMID: 7309855
NITR PACI	*Nitratireductor pacificus*	marine, denitrification	–	–	DOI: 10.1099/ijs.0.024356-0
NITR SP01	*Nitrincola sp. AK23*	aquatic, alkaline	–	–	DOI: 10.1016/j.syapm.2015.09.002
NOCA CONC	*Nocardia concava*	–	human pathogen	–	DOI: 10.1099/ijs.0.63280-0
NOCA TRAN	*Nocardia transvalensis*	–	human pathogen	–	DOI: 10.4103/0970-2113.99121
OLEI SOLI	*Oleiagrimonas soli*	soil, contaminant degradation	–	–	DOI: 10.1099/ijs.0.000158
PECT CARO	*Pectobacterium carotovorum*	soil, aquatic, aerosol, nitrogen fixation	insect endogastric flora, secretes antibiotics, plant pathogen	–	DOI: 10.5423/PPJ.OA.12.2013.0117
PHOT LUMI	*Photorhabdus luminescens*	–	endosymbiont of entomopathogenic nematode, insect pathogen, used as toxin against the host insect by the nematode	genomic	DOI: 10.1128/AEM.69.4.1890-1897.2003 10.1099/00207713-49-4-1645 microbewiki.kenyon.edu Search: Photorhabdus luminescens
PHOT TEMP	*Photorhabdus temperate*	–	endosymbiont of entomopathogenic nematode, insect pathogen, used as toxin against the host insect by the nematode	–	DOI: 10.1128/genomeA.01273-14 10.1099/ijs.0.2008/000273-0 10.1099/00207713-49-4-1645
PSEU FLUO	*Pseudomonas fluorescens*	soil, aquatic, nitrogen reducing	fungal pathogen, endosymbiont or endoparasite of amoebae, antifungal plant root symbiont, secretes antibiotics	–	DOI: 10.1128/CMR.00044-14 microbewiki.kenyon.edu Search: Pseudomonas fluroscens
PSEU MONT	*Pseudomonas monteilii*	soil, aquatic, nitrogen reducing	fungal pathogen, endosymbiont or endoparasite of amoebae, antifungal plant root symbiont, secretes antibiotics	–	DOI: 10.1099/00207713-47-3-846

PSEU VERO	*Pseudomonas veronii*	soil, aquatic, nitrogen reducing	fungal pathogen, endosymbiont or endoparasite of amoebae, antifungal plant root symbiont, secretes antibiotics	–	DOI: 10.1128/genomeA.00258-13 10.1186/s40793-016-0198-y
RAL SOLA	*Ralstonia solanacearum*	soil, voluntary dormancy	plant pathogen	–	DOI: 10.1111/mpp.12038
RALS EUTR	*Ralstonia eutropha* (a.k.a. *Cupriavidus necator*)	soil, aquatic, contaminant degradation	–	genomic	microbewiki.kenyon.edu Search: Ralstonia eutropha
RALS PICK	*Ralstonia pickettii*	wet soil, aquatic sediment	biofilm, opportunistic human pathogen	genomic	DOI: 10.1016/j.jhin.2005.08.015 microbewiki.kenyon.edu Search: Ralstonia pickettii
RALS SP01	*Ralstonia sp. UNC404CL21Col*	soil	plant hosted, forms communities	genomic	BioProject Accession: PRJNA213749 BioSample Accession: SAMN02743945
RHIZ ETLI	*Rhizobium etli*	soil, nitrogen fixation	legume root nodular symbiont	–	DOI: 10.1128/AEM.69.2.884-893.2003
RHIZ LEGU	*Rhizobium leguminosarum*	soil, nitrogen fixation	legume root nodular symbiont	plasmid	DOI: 10.1111/j.1365-294X.2004.02259.x
RHIZ PHAS	*Rhizobium phaseoli*	soil, nitrogen fixation	legume root nodular symbiont	–	PMID: 3584072
rhizavidin	*Rhizobium etli*	soil, nitrogen fixation	legume root nodular symbiont	plasmid	DOI: 10.1042/BJ20070076
RHOD PALU	*Rhodopseudomonas palustris*	wet soil, aquatic sediment, nitrogen fixation, carbon fixation, photosynthetic	–	plasmid	microbewiki.kenyon.edu Search: Rhodopseudomonas palustris
RHOD SP01	*Rhodanobacter sp. OR444*	soil, heavy metal resistance and purification	–	genomic	DOI: 10.1128/mBio.02234-15
SACC MARI	*Saccharomonospora marina*	marine sediment	–	–	DOI: 10.1099/ijs.0.017038-0 10.4056/sigs.2655905
SHEW DENI	*Shewanella denitrificans*	marine, denitrification	species in same genus participate sponge microflora	–	DOI: 10.1099/00207713-52-6-2211
shwanavidin	*Shewanella denitrificans*	marine, denitrification	species in same genus participate sponge microflora	genomic	DOI: 10.1074/jbc.M112.357186
STRE AVID	*Streptomyces avidinii*	soil, sporulation	secretes antibiotics	genomic	microbewiki.kenyon.edu Genus: Streptomyces
STRE FLAV	*Streptomyces flavotricini*	soil, sporulation	secretes antibiotics and antifungal agents	genomic	DOI: 10.1038/ja.2011.12 microbewiki.kenyon.edu Genus: Streptomyces
STRE KATR	*Streptomyces katrae*	soil, sporulation	secretes antibiotics	–	microbewiki.kenyon.edu Genus: Streptomyces
STRE VENE	*Streptomyces venezuelae*	soil, sporulation	secretes antibiotics	–	
STRE VIRG	*Streptomyces virginiae*	soil, sporulation	secretes antibiotics and antifungal agents	genomic	DOI: 10.1099/00221287-136-3-581 microbewiki.kenyon.edu Genus: Streptomyces

streptavidin	*Streptomyces violaceus* (former *S. avidinii*)	soil, sporulation	secretes antibiotics	–	microbewiki.kenyon.edu Genus: Streptomyces
strongavidin	*Strongylocentrotus purpuratus*	sea urchin	–	genomic	
tamavidin1	*Pleurotus cornucopiae*	wood decay fungus	weakly parasitic, nematophagous	–	
tamavidin2	*Pleurotus cornucopiae*	wood decay fungus	weakly parasitic, nematophagous	–	
VIBR GENO	*Vibrio genomosp.*	marine, bacterioplankton	secretes antibiotics	–	BioProject Accession: PRJNA164825
XANT ALBI	*Xanthomonas albilineans*	soil, aquatic	plant pathogen, opportunistic animal pathogen	–	PMID: 20572987
XANT AXON	*Xanthomonas axonopodis* (former *X. cassavae*)	soil, aquatic	plant pathogen, opportunistic animal pathogen	genomic	DOI: 10.1128/AEM.05189-11 micorebwiki.kenyon.edu Search: Xanthomonas axonopodis
XANT CAMP	*Xanthomonas campestris*	soil, aquatic	plant pathogen, opportunistic animal pathogen	genomic	DOI: 10.1094/PHYTO.2001.91.5.492 micorebwiki.kenyon.edu Search: Xanthomonas campestris
XANT CASS	*Xanthomonas axonopodis* (former *X. cassavae*)	soil, aquatic	plant pathogen, opportunistic animal pathogen	genomic	DOI: 10.1128/AEM.05189-11 micorebwiki.kenyon.edu Search: Xanthomonas axonopodis
XANT FUSC	*Xanthomonas Fuscans*	soil, aquatic	plant pathogen, opportunistic animal pathogen	genomic	DOI: 10.1128/AEM.05189-11 10.1186/1471-2164-14-761
XANT ORYZ	*Xanthomonas oryzae*	soil, aquatic	plant pathogen, opportunistic animal pathogen	genomic	DOI: 10.1111/j.1364-3703.2006.00344.x
XANT TRAN	*Xanthomonas translucens*	soil, aquatic	plant pathogen, opportunistic animal pathogen	genomic	DOI: 10.1094/PHYTO-08-16-0286-R
XANT VASI	*Xanthomonas vasicola*	soil, aquatic	plant pathogen, opportunistic animal pathogen	genomic	DOI: 10.1111/j.1365-3059.2009.02124.x
xenavidin	*Xenopus tropicalis*	clawed frog	–	genomic	
zebavidin	*Danio rerio*	freshwater fish	–	–	

Appendix B. Web-user-interface images of the DATAvidin database.

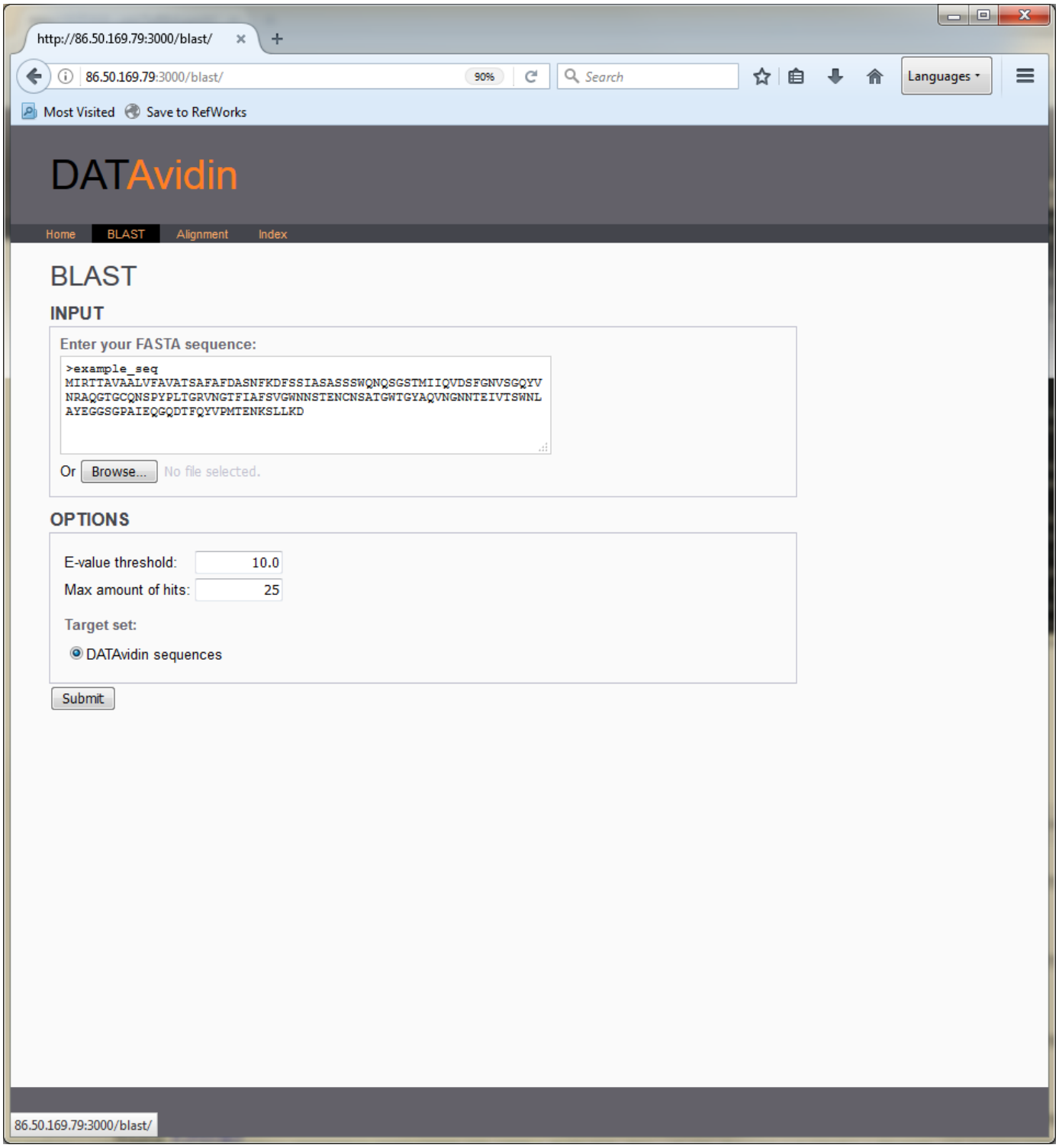


Figure Appendix B1. Blast input –view from the web-based interface for DATAVidin-database.
The desired fasta-sequence(s) can be input either via a fasta-formatted text-file or as raw input to input field. The BLAST+-query supports at the moment only two additional parameters: E-value threshold, which sets a minimum value for the query hit quality; and the maximum amount of hits shown in results. Additionally, there can later option to use other external databases, such as NCBI GenBank, but at the moment the option defaults in searching from DATAvidin-database and cannot be changed.

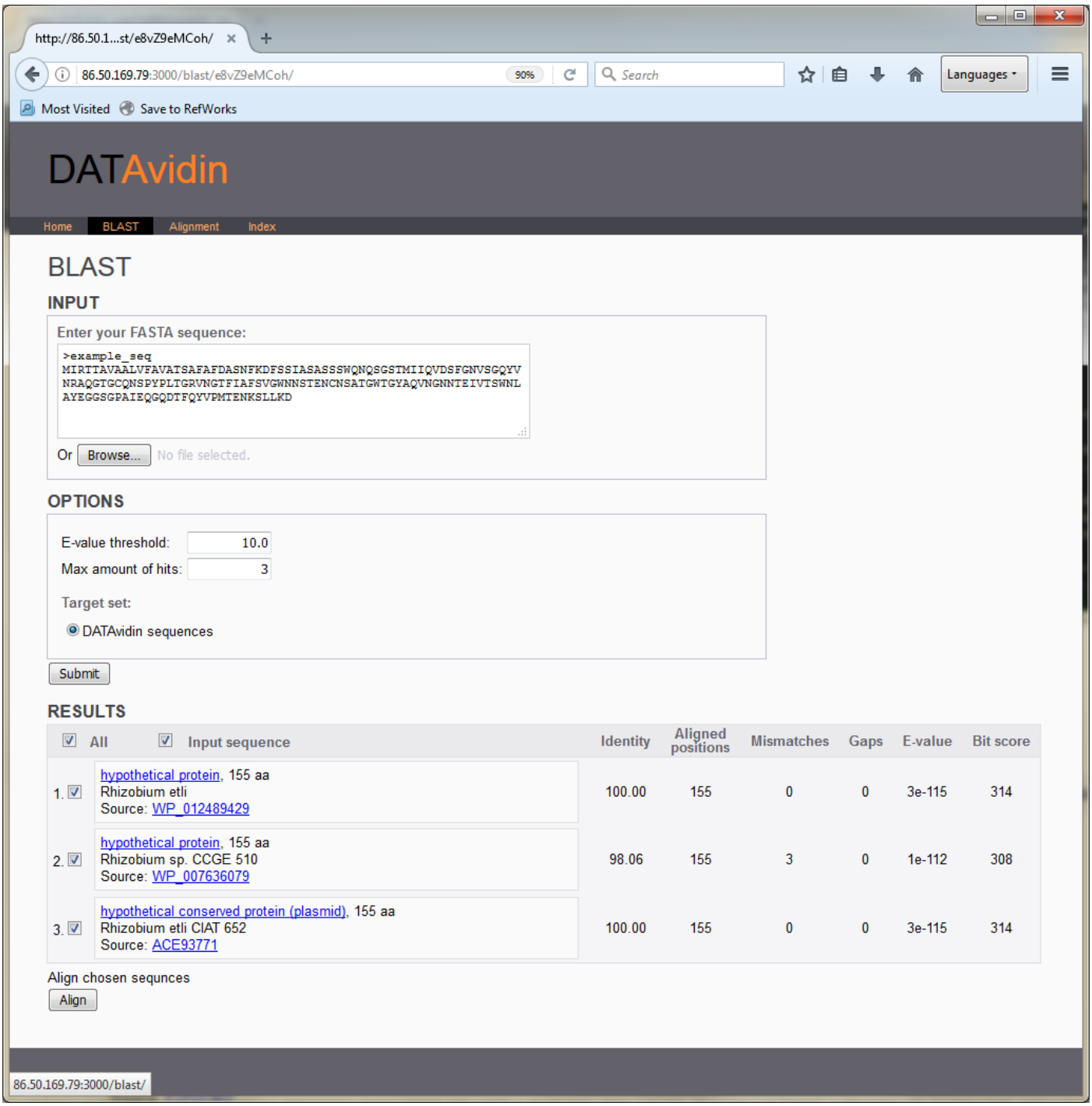


Figure Appendix B2. Blast results –view from the web-based interface for DATAVidin-database.
The results are shown in a table format with tick-boxes to choose interesting sequences for alignment. In addition to the results, the input-sequence can be chosen to be included in alignment too. The result-view includes links to both hit corresponding GenBank entries and sequence page within DATAvidin (Figure Appendix B6.).

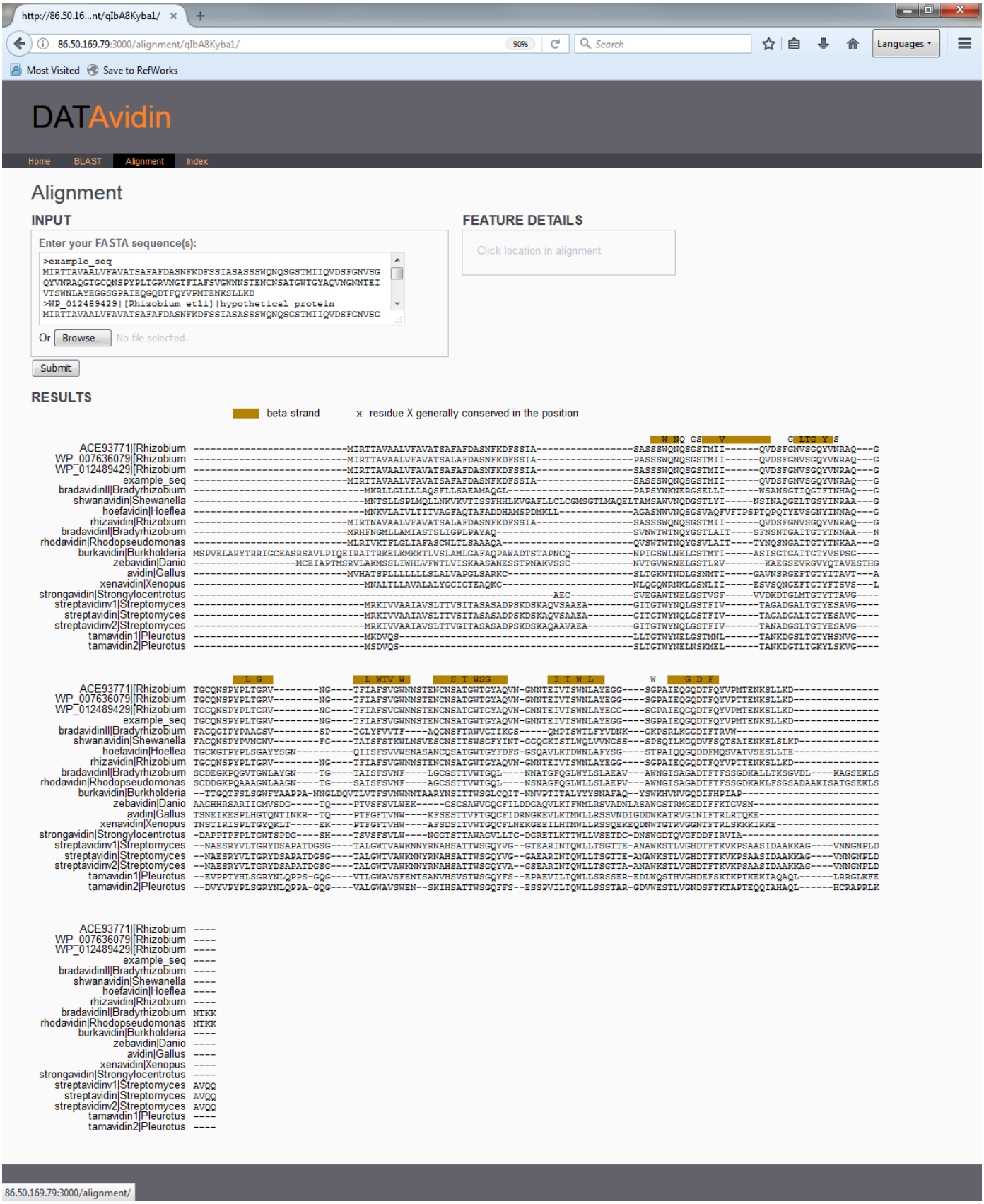


Figure Appendix B3. Alignment results –view from the web-based interface for DATAVidin-database.
The alignment can be run with either direct input, again as fasta-formatted text-file or raw input, or from the selected BLAST query results (Figure Appendix B2). The alignment is shown in 160 AA segments and the positions of streptavidin β-strands and conserved amino acids are marked above the alignment. When such location in the alignment is clicked, the feature detail box will show details of the β-strands and conserved amino acids at the position (Figure Appendix B4).

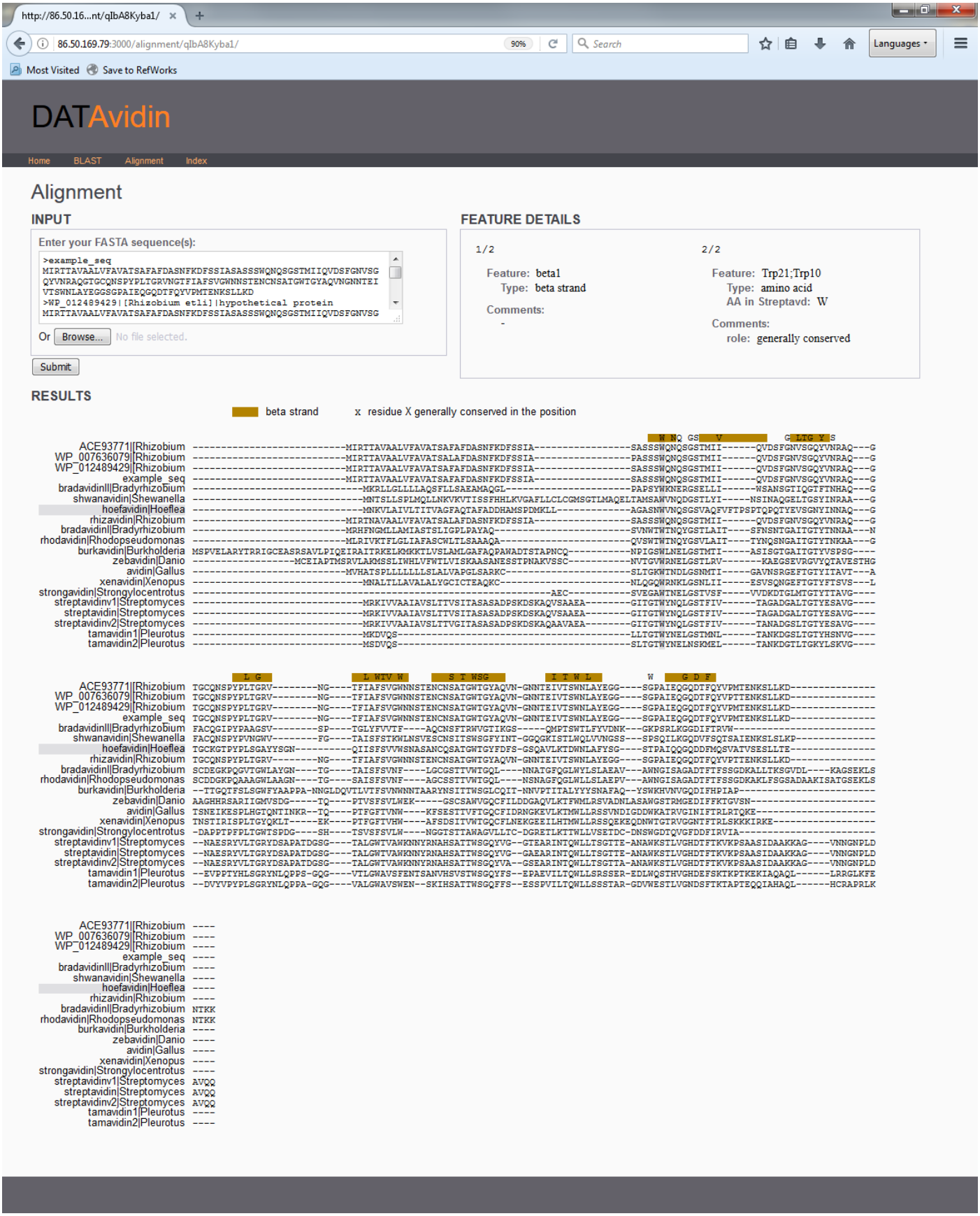


Figure Appendix B4. Alignment results –view with an alignment position clicked from the web-based interface for DATAVidin-database.

The alignment results –view (Figure Appendix B4) highlights the position in alignment upon mouse-hover and shows location specific details upon click in the Feature details -box.

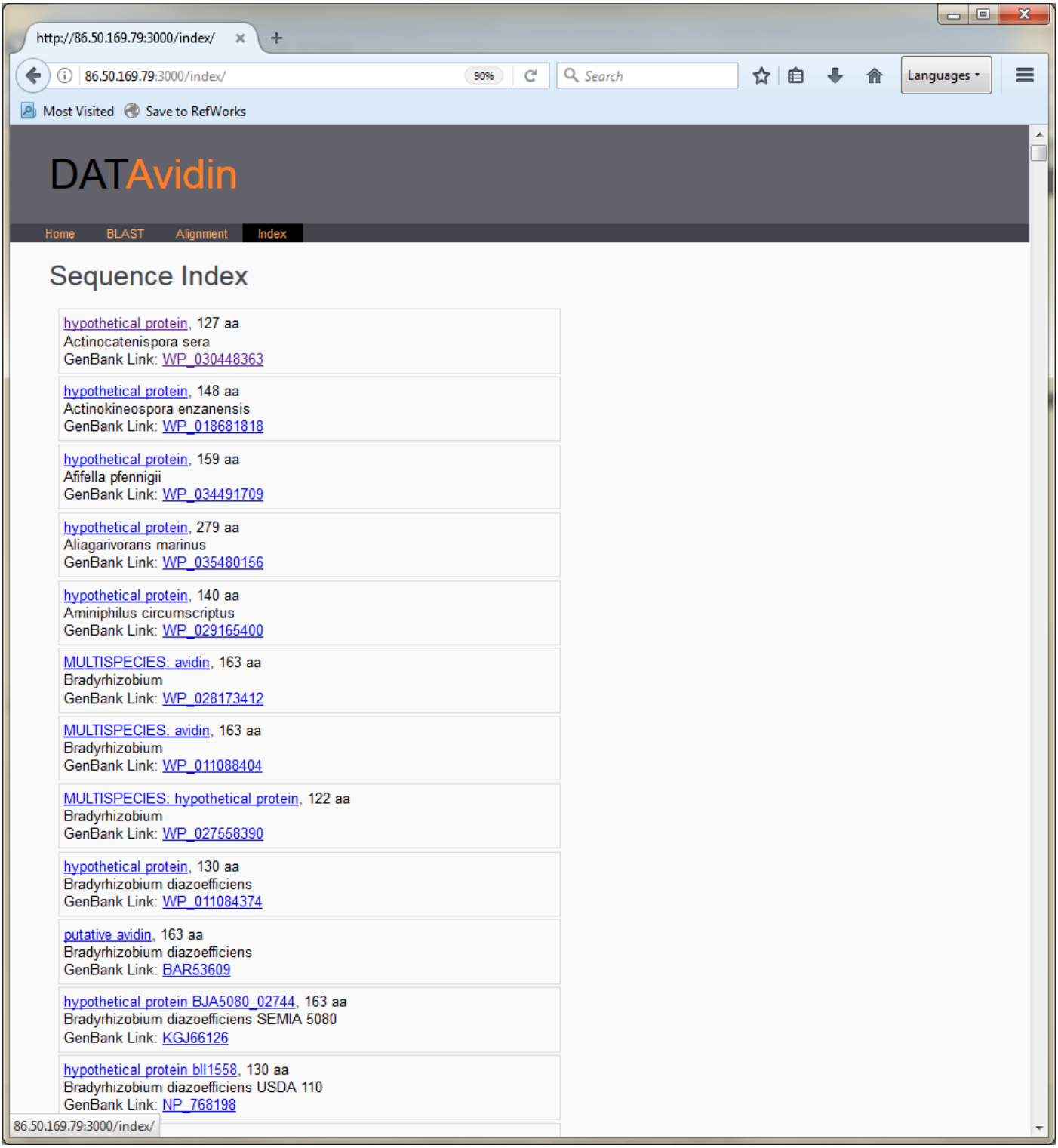


Figure Appendix B5. Sequence Index –view from the web-based interface for DATAVidin-database.
The sequence index –view lists the DATAvidin-database protein sequences with similar information displayed as in BLAST-query results (Figure Appendix B2). The view gives two links for each database entry: a link to GenBank source and a link to sequence page within DATAvidin (Figure Appendix B6.).

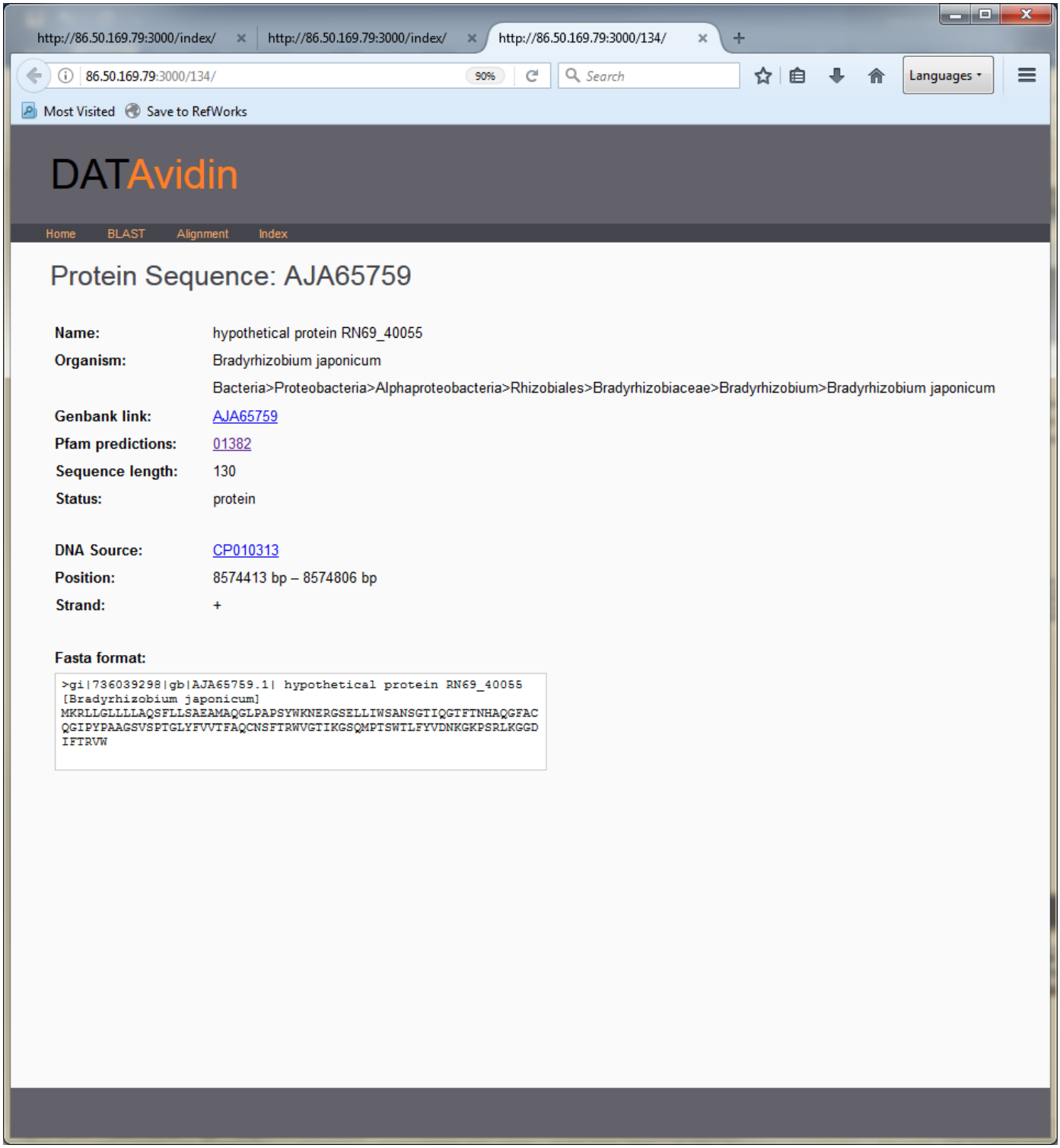


Figure Appendix B6. Sequence –view from the web-based interface for DATAVidin-database.
The sequence –view shows most of the key information stored in DATAvidin-database per requested sequence. The information contains links for corresponding GenBank entry and PFAM site for the PFAM protein family predictions. The DNA source and other DNA specific information will be displayed only if the the sequence has a corresponding DNA entry in DATAvidin-database.

Appendix C. The enrichment analysis results for genes in avidin proximity.

Table Appendix C. Enrichment analysis results per Gene Ontology-term and related function. The light grey rows highlight the DNA processing and mobile elements related pathways, medium grey the key metabolic pathways, and dark grey the defence mechanism pathways. The GO-functions are ordered from smallest P-value to highest and the 0,05 and 0,1 are marked in the table with heavier line between the rows.

GO-term	N_{avid}	N_{total}	Odds ratio	P-value	GO-function
GO:0047632	1	2	562.17	0.0027	agmatine deiminase activity
GO:0005540	1	3	374.78	0.0036	hyaluronic acid binding
GO:0000150	3	343	9.95	0.0038	recombinase activity
GO:0003933	1	8	140.54	0.0080	GTP cyclohydrolase activity
GO:0009446	1	8	140.54	0.0080	putrescine biosynthetic process
GO:0000156	2	152	14.88	0.0086	phosphorelay response regulator activity
GO:0004474	1	9	124.92	0.0089	malate synthase activity
GO:0006097	1	9	124.92	0.0089	glyoxylate cycle
GO:0004668	1	10	112.43	0.0098	protein-arginine deiminase activity
GO:0006323	1	10	112.43	0.0098	DNA packaging
GO:0019217	1	10	112.43	0.0098	regulation of fatty acid metabolic process
GO:0004803	3	527	6.47	0.0123	transposase activity
GO:0000062	1	14	80.30	0.0133	fatty-acyl-CoA binding
GO:0000334	1	14	80.30	0.0133	3-hydroxyanthranilate 3,4-dioxygenase activity
GO:0006313	3	545	6.26	0.0134	transposition, DNA-mediated
GO:0016779	2	220	10.28	0.0172	nucleotidyltransferase activity
GO:0051607	1	20	56.21	0.0186	defense response to virus
GO:0004565	1	27	41.64	0.0248	β-galactosidase activity
GO:0030145	1	27	41.64	0.0248	manganese ion binding
GO:0004497	1	31	36.26	0.0282	monooxygenase activity
GO:0006858	1	31	36.26	0.0282	extracellular transport
GO:0010309	1	31	36.26	0.0282	acireductone dioxygenase [iron(II)-requiring] activity
GO:0006355	10	5323	2.18	0.0300	regulation of transcription, DNA-templated
GO:0043571	1	33	34.06	0.0300	maintenance of CRISPR repeat elements
GO:0019012	1	36	31.23	0.0326	virion
GO:0006310	3	852	3.99	0.0420	DNA recombination
GO:0005976	1	50	22.48	0.0446	polysaccharide metabolic process
GO:0005506	2	417	5.42	0.0547	iron ion binding
GO:0050518	1	62	18.13	0.0548	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase activity
GO:0004177	1	65	17.29	0.0574	aminopeptidase activity
GO:0004499	1	73	15.40	0.0641	N,N-dimethylaniline monooxygenase activity
GO:0016846	1	76	14.79	0.0666	carbon-sulfur lyase activity
GO:0020037	2	552	4.09	0.0886	heme binding
GO:0008299	1	105	10.70	0.0906	isoprenoid biosynthetic process
GO:0009975	1	107	10.50	0.0922	cyclase activity
GO:0045735	1	107	10.50	0.0922	N2-acetyl-L-aminoadipate semialdehyde dehydrogenase activity
GO:0004519	1	109	10.31	0.0938	endonuclease activity

GO:0009966	1	111	10.12	0.0954	regulation of signal transduction
GO:0000160	3	1274	2.66	0.1075	phosphorelay signal transduction system
GO:0006099	1	127	8.85	0.1083	tricarboxylic acid cycle
GO:0005507	1	140	8.02	0.1186	copper ion binding
GO:0003700	6	3396	2.01	0.1322	transcription factor activity, sequence-specific DNA binding
GO:0007155	1	162	6.93	0.1358	cell adhesion
GO:0006118	1	167	6.73	0.1397	obsolete electron transport
GO:0045892	1	182	6.17	0.1512	negative regulation of transcription, DNA-templated
GO:0009055	2	796	2.83	0.1604	electron carrier activity
GO:0004872	1	200	5.61	0.1648	receptor activity
GO:0016705	1	202	5.56	0.1663	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0006950	1	209	5.37	0.1715	response to stress
GO:0098519	1	209	5.37	0.1715	nucleotide phosphatase activity, acting on free nucleotides
GO:0016987	2	842	2.68	0.1749	sigma factor activity
GO:0006352	2	846	2.66	0.1762	DNA-templated transcription, initiation
GO:0005524	1	4248	0.26	0.1922	ATP binding
GO:0009236	1	260	4.32	0.2085	cobalamin biosynthetic process
GO:0050661	1	282	3.98	0.2240	NADP binding
GO:0051287	2	1029	2.19	0.2357	NAD binding
GO:0043115	1	306	3.67	0.2405	precorrin-2 dehydrogenase activity
GO:0019354	1	308	3.64	0.2418	siroheme biosynthetic process
GO:0009058	2	1062	2.12	0.2465	biosynthetic process
GO:0055085	3	1925	1.76	0.2489	transmembrane transport
GO:0003824	1	3930	0.28	0.2729	catalytic activity
GO:0006779	1	359	3.12	0.2758	porphyrin-containing compound biosynthetic process
GO:0016740	1	378	2.97	0.2880	transferase activity
GO:0015074	1	383	2.93	0.2912	DNA integration
GO:0022857	1	388	2.89	0.2944	transmembrane transporter activity
GO:0008236	1	420	2.67	0.3143	serine-type peptidase activity
GO:0006813	1	442	2.54	0.3277	potassium ion transport
GO:0006810	1	3596	0.31	0.3865	transport
GO:0005515	2	1550	1.45	0.4050	protein binding
GO:0016747	1	581	1.93	0.4066	transferase activity, transferring acyl groups other than amino-acyl groups
GO:0043565	2	1561	1.44	0.4084	sequence-specific DNA binding
GO:0016301	1	600	1.87	0.4167	kinase activity
GO:0071949	1	650	1.72	0.4423	FAD binding
GO:0008033	1	691	1.62	0.4625	tRNA processing
GO:0000155	1	742	1.51	0.4866	phosphorelay sensor kinase activity
GO:0003677	6	5421	1.25	0.4911	DNA binding
GO:0003676	1	788	1.42	0.5074	nucleic acid binding
GO:0046872	1	817	1.37	0.5201	metal ion binding
GO:0004871	1	850	1.32	0.5341	signal transducer activity

GO:0005975	1	938	1.19	0.5696	carbohydrate metabolic process
GO:0016491	4	3725	1.20	0.5795	oxidoreductase activity
GO:0008152	2	4090	0.54	0.5927	metabolic process
GO:0016021	2	4277	0.52	0.5975	integral component of membrane
GO:0050660	1	1077	1.04	0.6202	flavin adenine dinucleotide binding
GO:0007165	1	1080	1.03	0.6213	signal transduction
GO:0016020	3	5033	0.66	0.6339	membrane
GO:0016787	2	1881	1.19	0.6869	hydrolase activity
GO:0055114	6	6112	1.10	0.8252	oxidation-reduction process
GO:0005215	1	1948	0.57	1.0000	transporter activity
GO:0006508	1	1288	0.87	1.0000	proteolysis