EVOLUTION OF PROTEINS INVOLVED IN RESPONSE TO REACTIVE OXYGEN SPECIES (=ROS)

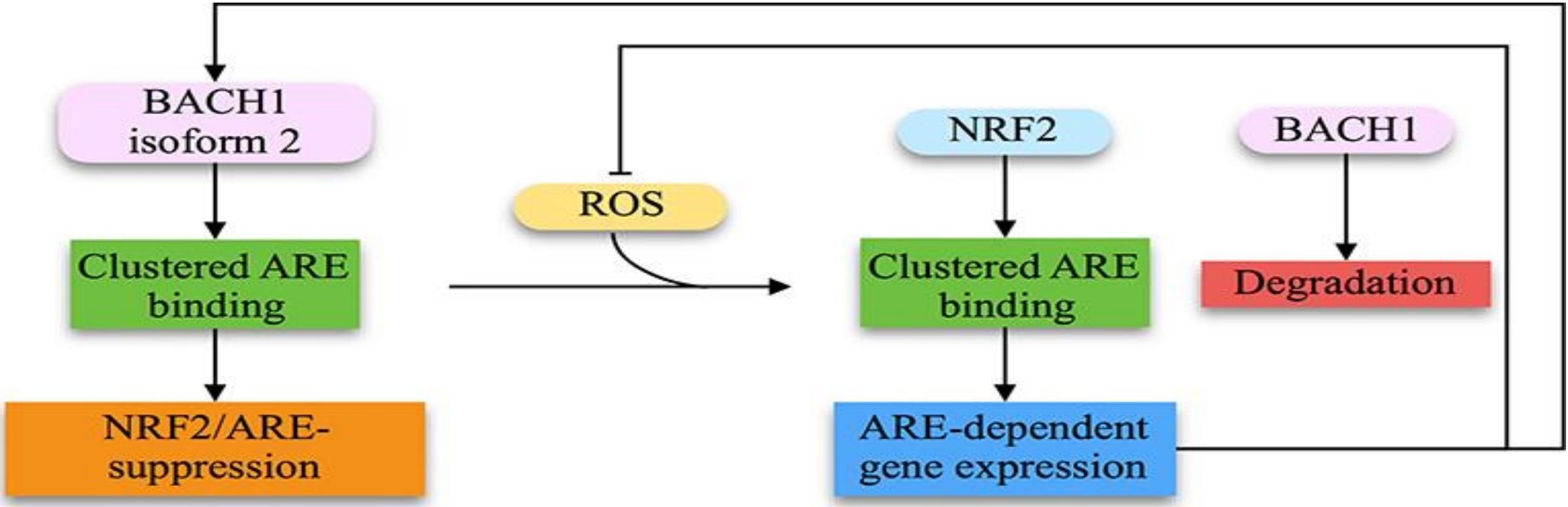
Vassily Lyubetsky (http://lab6.iitp.ru),
Oleg Zverkov,
Gregory Shilovsky,
Lev Rubanov,
Alexandr Seliverstov

Institute for Information Transmission Problems (IITP RAS)
Faculty of Mechanics and Mathematics &
Belozersky Institute of Physico-Chemical Biology (MSU)

Moscow, Russia 2020

In many vertebrates, **Nrf2** (Nfe2l2) and Bach proteins (**Bach1** and **Bach2**), are components of the **regulatory network** (e.g., in *Mus musculus*) that is, among other things, associated with **regulation of the level of reactive oxygen species** (=**ROS**).

The **functioning** of this network is related to **species-specific LIFESPAN**.



These proteins usually function by **binding with other proteins** ('typically **Maf**) and **DNA**.

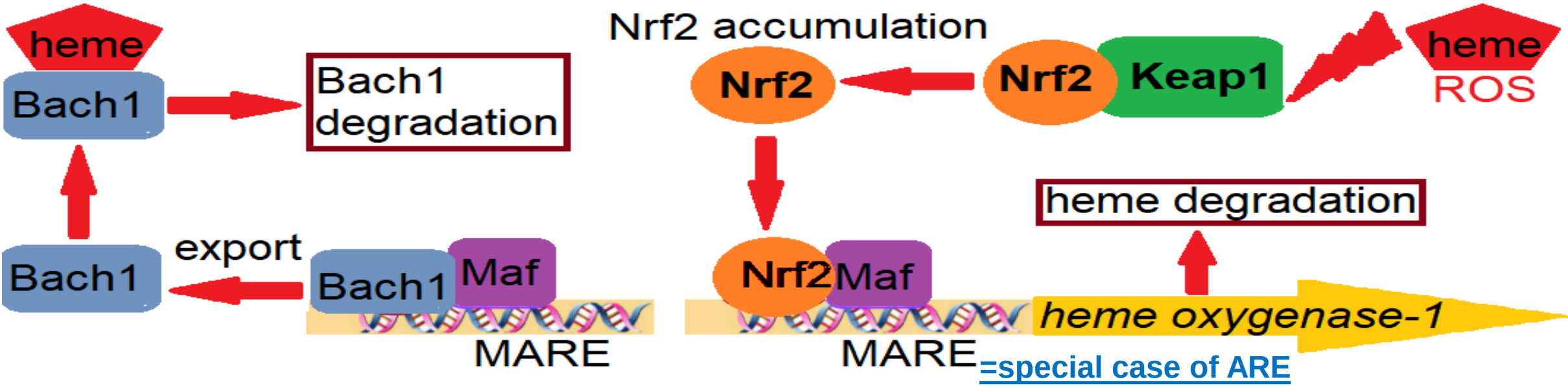
Nrf2 and **Bach1 are antagonistic** in **controlling ROS levels**,

which is also due to the **competition for MARE** sites.

(Analogously we investigated **Keap1**, **β-TrCP**, **c-Myc**, **GSK3b**, but now only concentrate on **Nrf2** and **Bach**.)

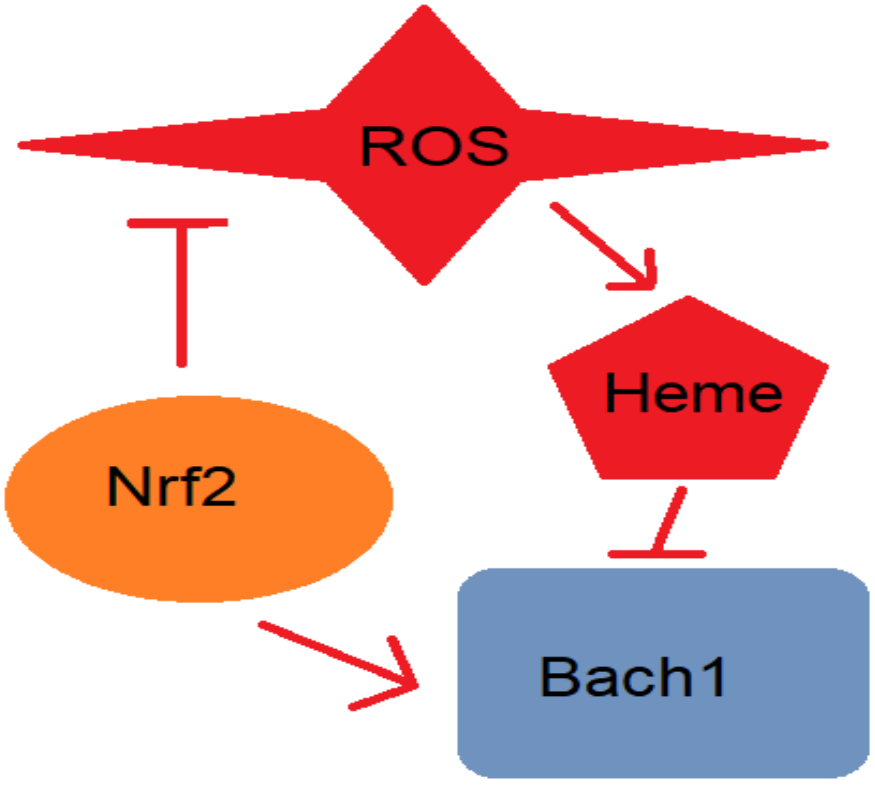
Let us EXEMPLIFY this function in detail.

Competition between Nrf2 and Bach1 on MARE in response to oxidative stimuli



In oxidative stress, Nrf2 **dissociates** from Keap1 & Nrf2 degradation is **inhibited**; so Nrf2 accumulates in the cytoplasm and **translocates** into the nucleus. Then, Nrf2 **binds** to MAREs as a heterodimer with small Mafs, which activates the **expression** of heme oxygenase-1 (and 200 other genes), thus **degrading heme**. At the same time, Bach1 is **displaced** from MAREs and **exported** out the nucleus. **IF insufficient heme** is available, Bach1 is **released** from heme and **again binds** Mafs and DNA.

Thus, the relationship between Nrf2 and Bach1 is critical for the **regulation** of heme oxygenase-1 expression, which **provides for heme degradation** according to this schematic diagram:



Now the small topic:

the heme-binding regions in Bach1 and Bach2 are **dissimilar**; in particular, they **differ** in the NUMBER of **regulatory cysteine-proline motifs critical for heme binding** (also called **CP-motifs**).

The **heme-binding motifs** (including those CP-motifs) have been identified by us in Bach **in chordates**;

as well as the **sites of Bach homodimerization**.

The **functioning** of this network is **related to aging** as well as to many **human diseases**.

E.g., mice homozygous for a **knock-out *Nrf2*** exhibit increased sensitivity to **oxidative stress** in a variety of tissues and cells (including the brain, liver, erythrocytes, and spleen), abnormal tooth **enamel**, and abnormal response to various '**injuries**, chemical **treatments**, and induced **inflammatory diseases**.

Single nucleotide variants in *Bach2* have been linked to a number of **autoimmune diseases** in humans. Mendelian *Bach2*-related immunodeficiency and autoimmunity **(BRIDA) syndrome** in humans is caused by haploinsufficiency of **this transcription factor** resulting from germline mutations.

We present the results of the computer-aided search for Nrf2 and Bach in deuterostomes. In other words, we found true orthologs of human proteins

NFE2L2 (NP_006155.2), **BACH1** (NP_001177.1), **BACH2** (NP_068585.1).

There is a well-known problem of similarity thresholds between two sequences as a function of the evolutionary distance between species.

Accordingly, we used criteria to **determine homologs with** "**the same molecular function**" in species **evolutionary distant** from human.

"**The same function**" means here the **ANALOGOUS RESPONSE TO OXIDATIVE STIMULI**.

Remember that:

bZIP domain, **typical of both Nrf2** and **Bach** proteins, is determined **even in Choanoflagellates**;

both bZIP and **BTB** domains are widespread **even in plant proteins**.

This shown how hard can be **determine Nrf2** and **Bach** in **distance species**. For that we used the following criteria.

Apart from the similarity extracted from the **local** and **global alignments** with the human genes, these **criteria include**: the presence or absence of the **BTB** and **bZIP-Maf domains**, satisfactory **alignment of the secondary structure**, high similarity of the **3D structures** in the N- and C-terminal regions of the genes; the **absence of "wrong" domains** like kelch-type beta-propeller in kelch-like proteins, BTB in Bach (for Nrf2 case), bZIP-Jun in Jun, and other specific domains from bZIP proteins; as well as high 3D structure similarity with NFE2L2 at the C-terminus coupled with a significant difference from "wrong" bZIP proteins at the N-terminus.

Also the method **takes into consideration**: **CP-motifs**, **N-hooks** and presence of **specific amino acids**.

Additionally **the method** includes the **SPLITTING** of the unrooted tree **into the clades of Bach and Nrf2 separated with a nearly 100% support** (for dividing all found proteins into the **Nrf2** and **Bach1–2 groups**). The tree is shown below. Similar splitting has been done for other protein groups.

This exemplifies our METHOD to the problem of **mutual evolution of the Nrf2 and Bach1–2 genes**.

Among invertebrate deuterostomes, Bach proteins have been **found only in *Ciona intestinalis* and *C. savignyi***

(tunicates diverged early from other chordates). Shown below.

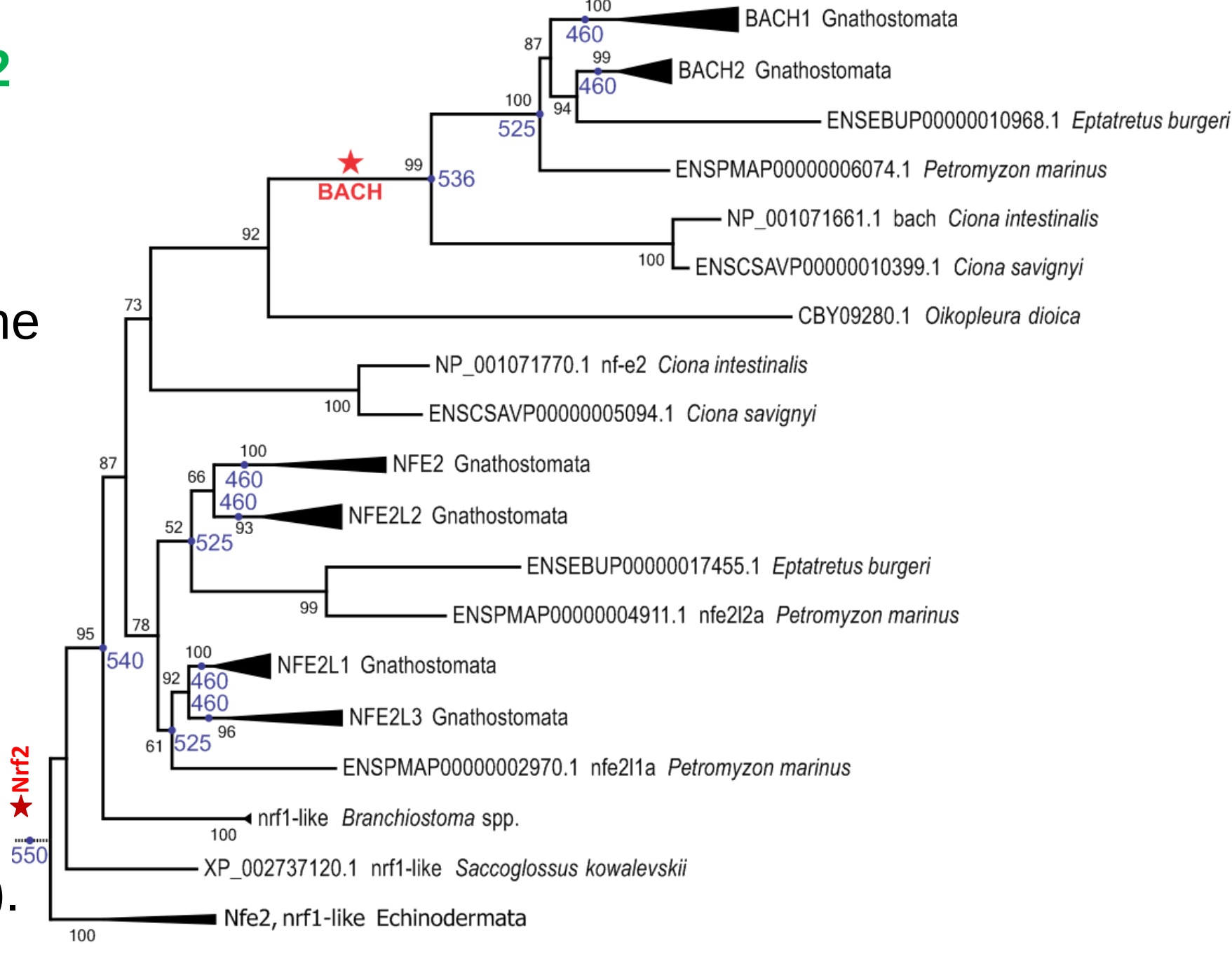
At least one Bach protein has been **found in each vertebrate**.

On these grounds, the **Bach** and **Nrf2** **protein tree** was generated:

The **Bach** and **Nrf2** **protein tree**

(blue indicates MYA)

Thus, the Bach gene emerges in the **Olfactores** (a subphylum within the Chordata that comprises the Tunicata and the Vertebrata).



Nevertheless, the **genomic rearrangement** that gave rise to the Bach genes remains unclear.

The *Bach* is most similar to *Nrf2* in the early diverging deuterostomes (by bZIP), and the BTB exists in dozens of ancient proteins (e.g. in the BTB-ZF family).

Therefore we think that ***Bach* emerged by duplication of *Nrf2***, an ortholog of *Nfe2l2*, in the **chordate ancestor**. The **ancestral Bach was provided by BTB from those ancient proteins**.

(Bach has to contain both BTB и bZIP.)

To follow **our METHOD**,

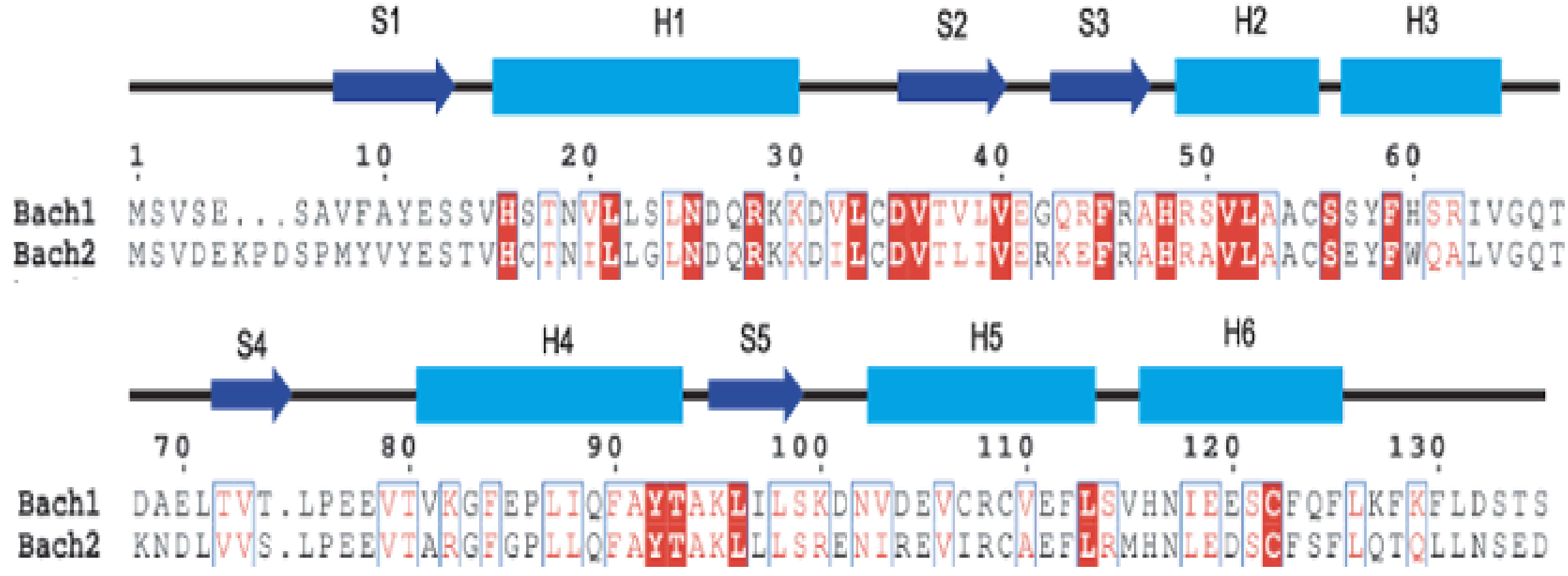
we **enumerates characteristics** of the **Nrf2** and **Bach** proteins

that we have used:

The **primary** and **secondary structures** of the BTB domain

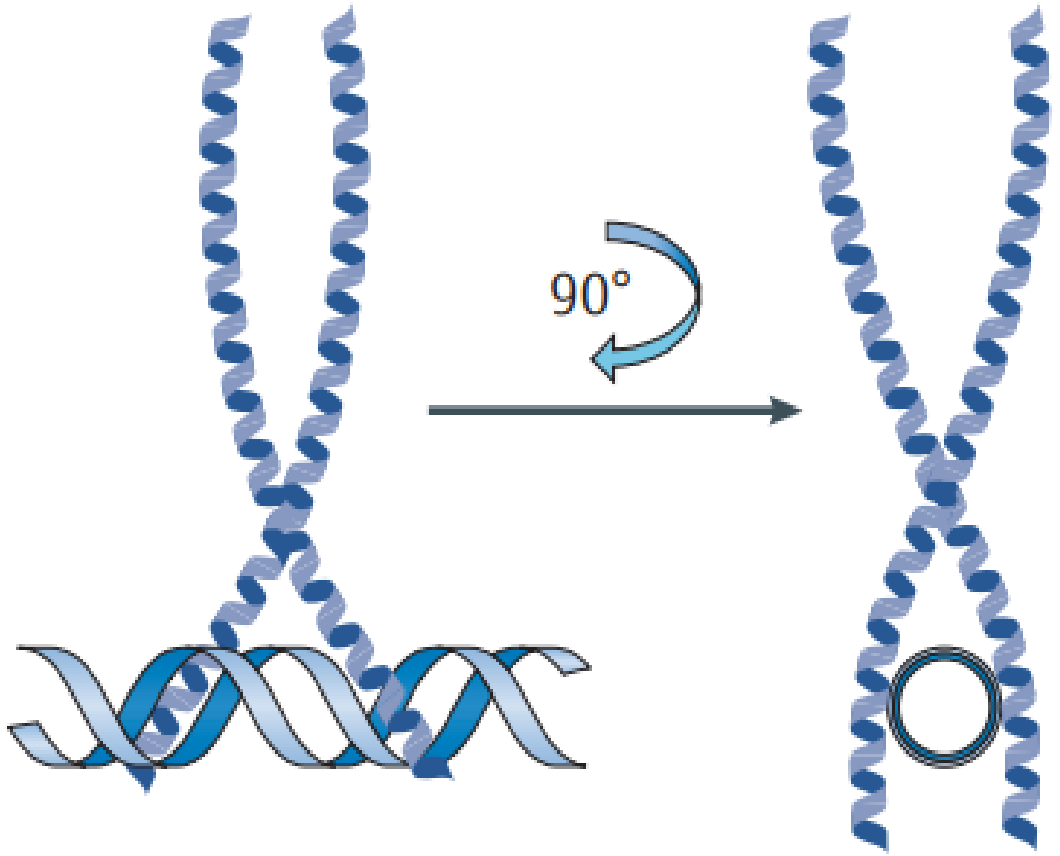
in mouse (in Bach1) and human (in Bach2);

S, beta-sheets; H, alpha-helices:



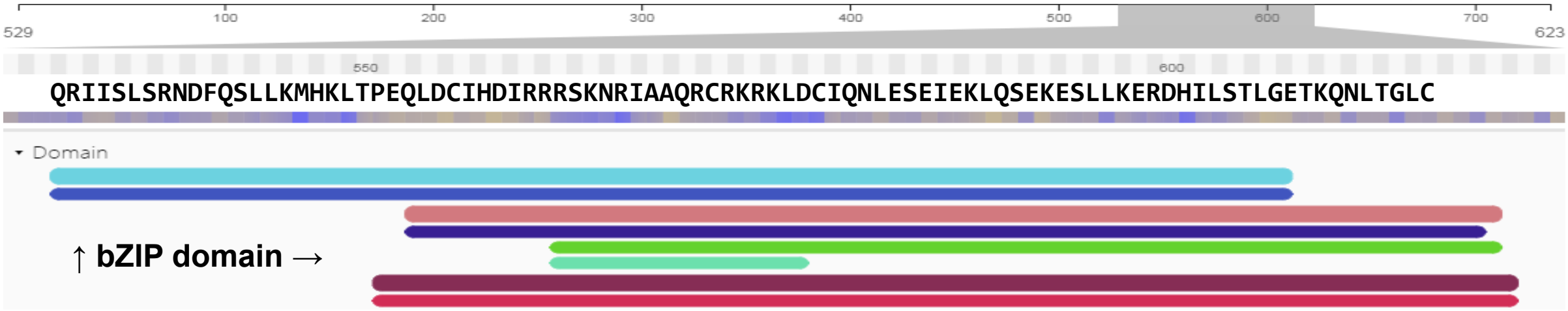
Dimeric interaction of **bZIP** at DNA in

Bach1 and **Nrf2** – both with Mafs:

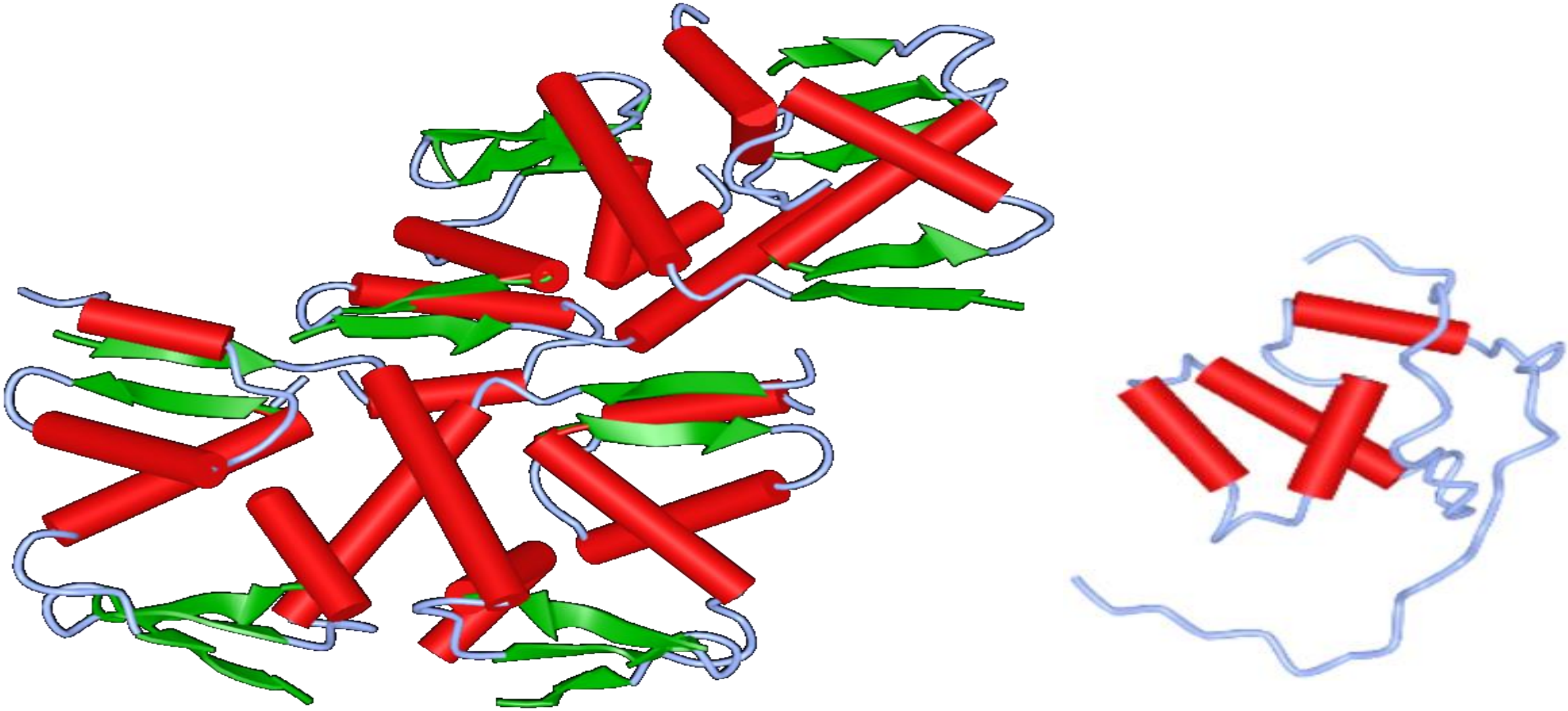


Their **primary** structure is shown

underneath:

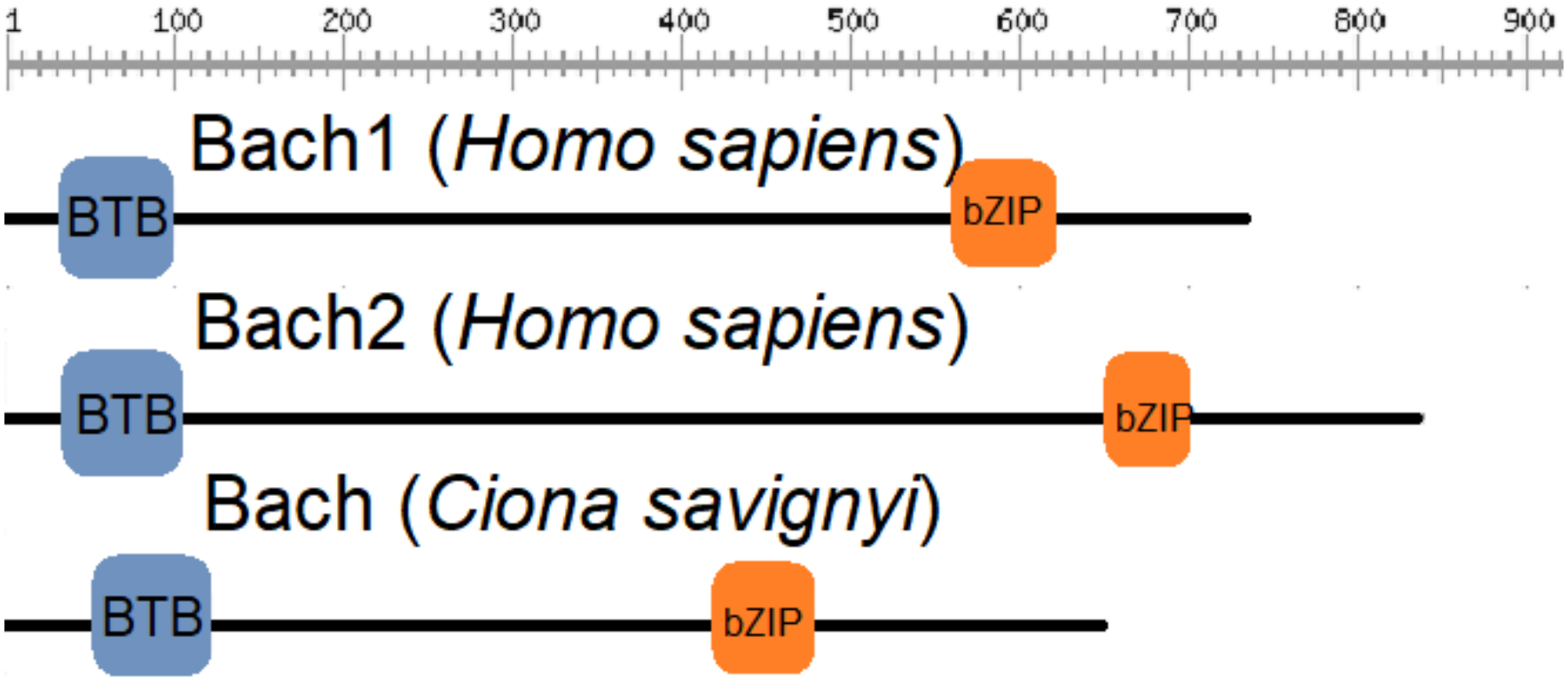


The tertiary structure of BTB domain (4 BTB-chains) **and the bZIP domain** (DNA binding fragment) **– in human**:

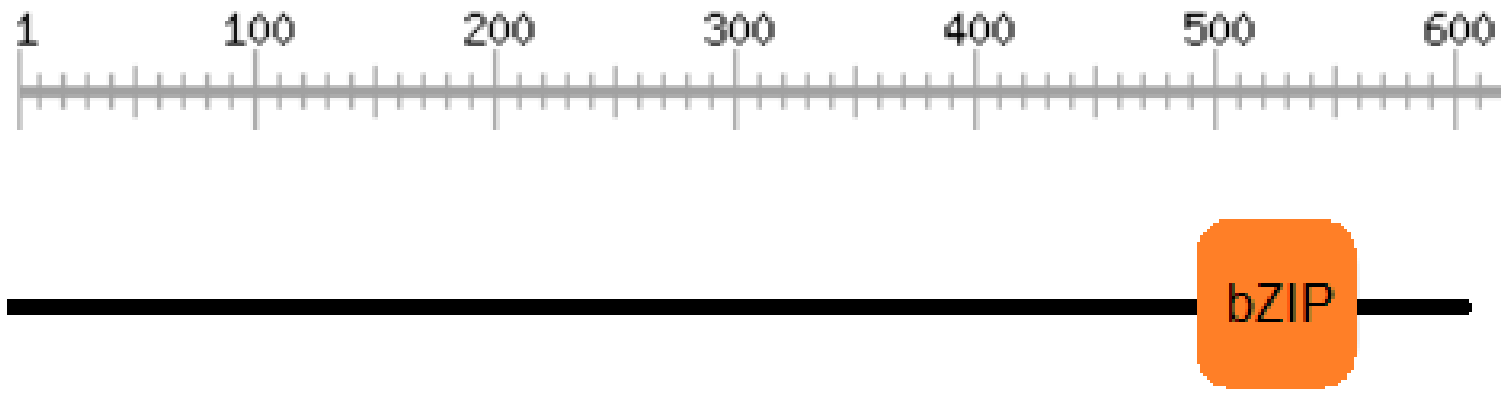


Conserved domains BTB and bZIP in Bach1–2:

BTB domain at the N-terminus and the bZIP domain at the C-terminus:



Conserved domain bZIP at the C-terminus in Nrf2:



(Nrf2 **includes no** BTB domain.)

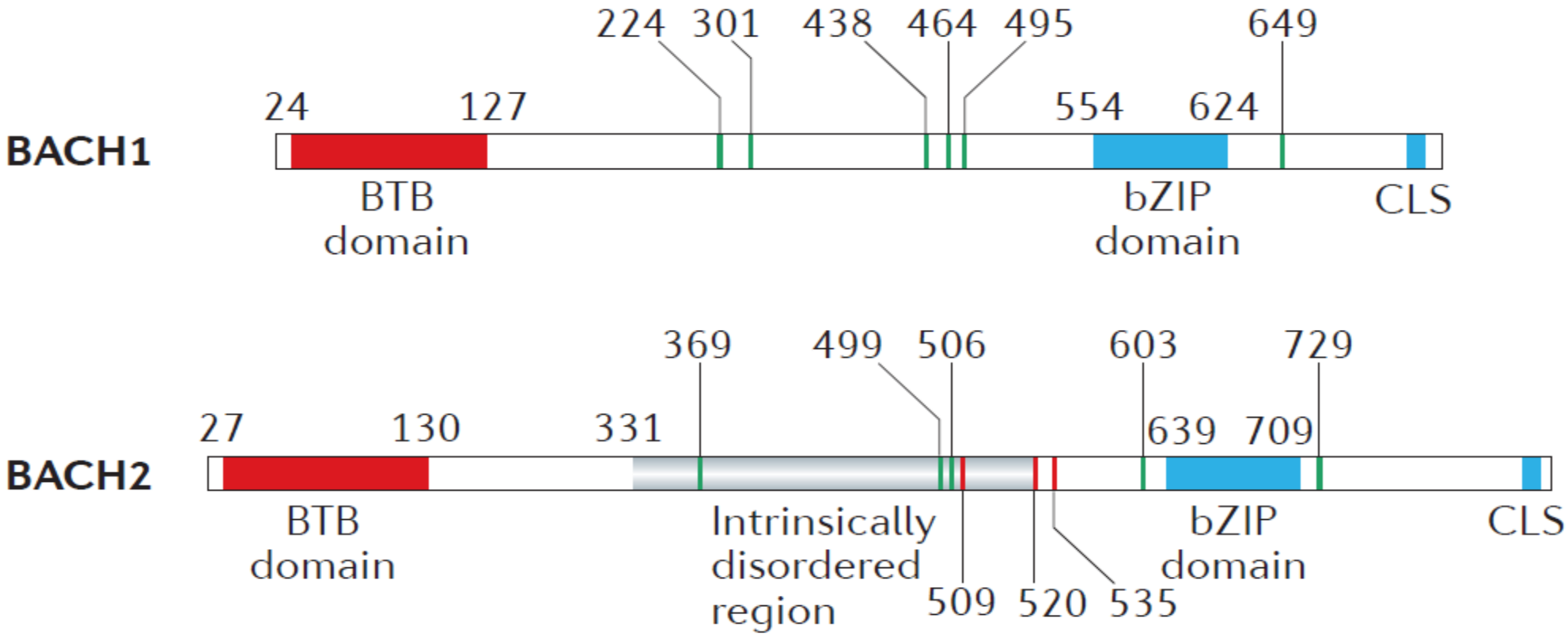
Thus, the **BTB domain** is 'present near the N-terminus in proteins that contain the Kelch motif.

(The BTB domain mediates homodimerization and in some instances heterodimerisation as well as it mediates transcriptional repression.)

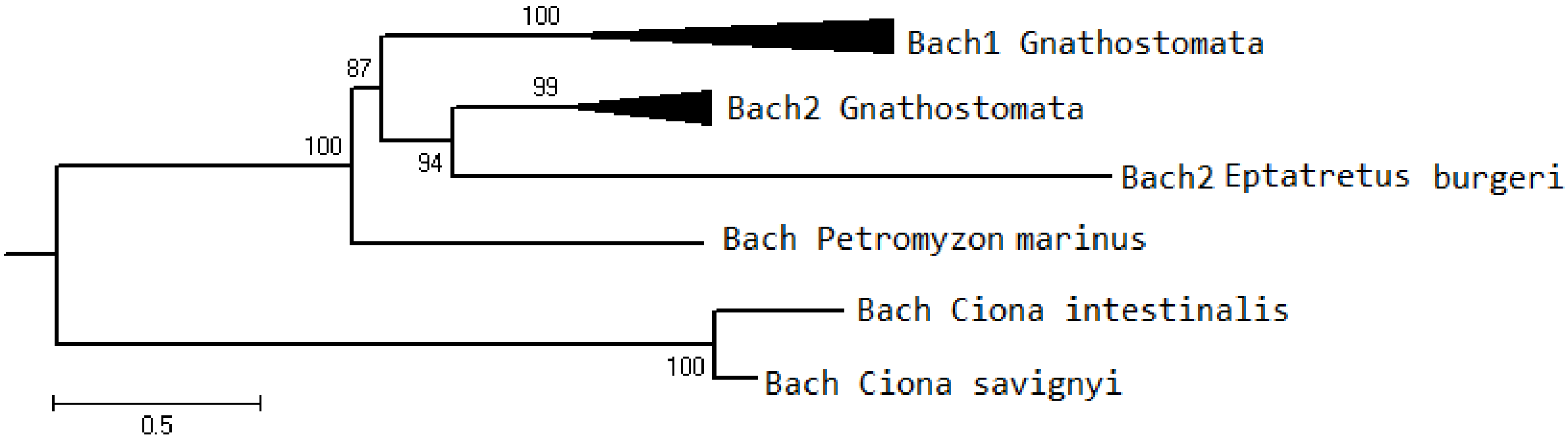
The Basic Leucine Zipper Domain-Maf-type (**bZIP-Maf domain**) is found in many DNA-binding proteins. <One part of the domain mediates sequence specific DNA binding, and the leucine zipper is required to hold together (dimerize) two DNA-binding regions. The DNA-binding region comprises a number of basic amino acids such as arginine and lysine.>

The motifs are termed the Maf-recognition element (**MARE**).

Detailed domain structures and post-translational modifications of the **mouse Bach1–2**. Green bars indicate **cysteine–proline residues**; numbers indicate the positions of cysteine. Red lines indicate selected identified phosphoresidues. CLS is the cytoplasmic localization signal:



Now we continiue: Bach1 and Bach2 proteins in Human are more similar to each other **than to any protein** in *Ciona* spp., and **we got the tree**:

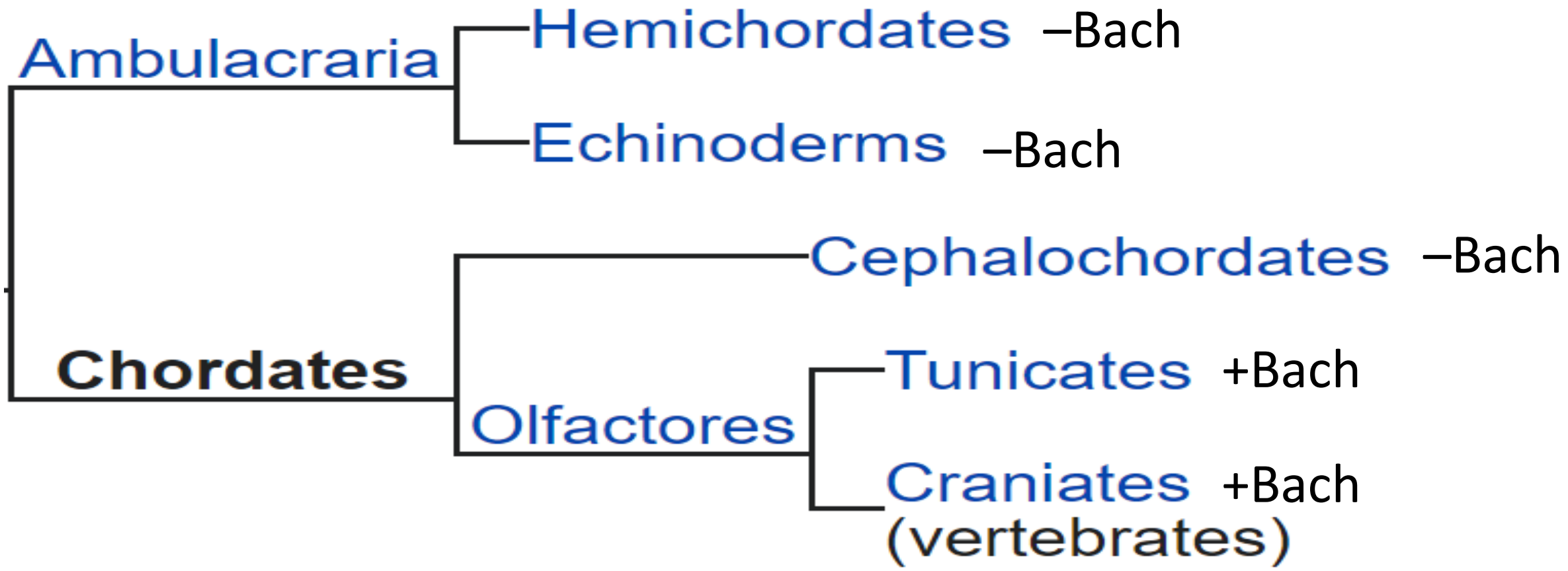


The following evolutionary assumptions have been proposed:

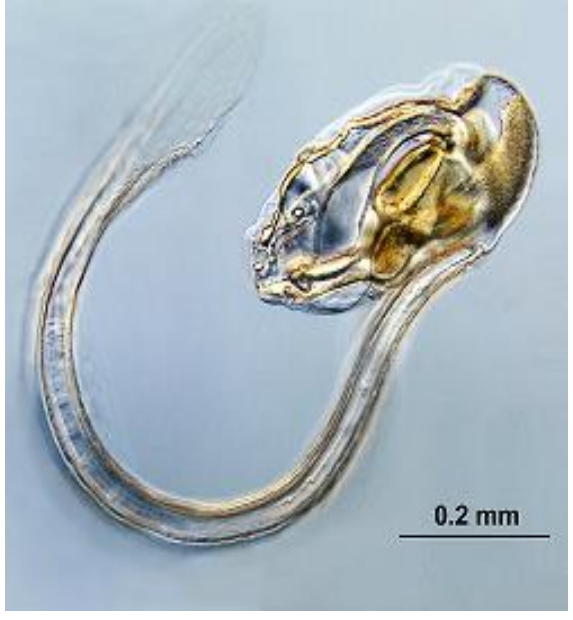
- the common ancestor of the *BACH* genes **existed in chordates before the divergence of vertebrates** since it exists in *Ciona* spp.;

- two *BACH* genes **emerged in gnathostomatous**.

The *Bach* genes are missing in lancelets, hemichordates, and echinoderms. The absence of *Bach* in *Branchiostoma* spp. (Cephalochordates) agrees with the proposed closer phylogenetic relation between vertebrates and tunicates rather than between vertebrates and lancelets.



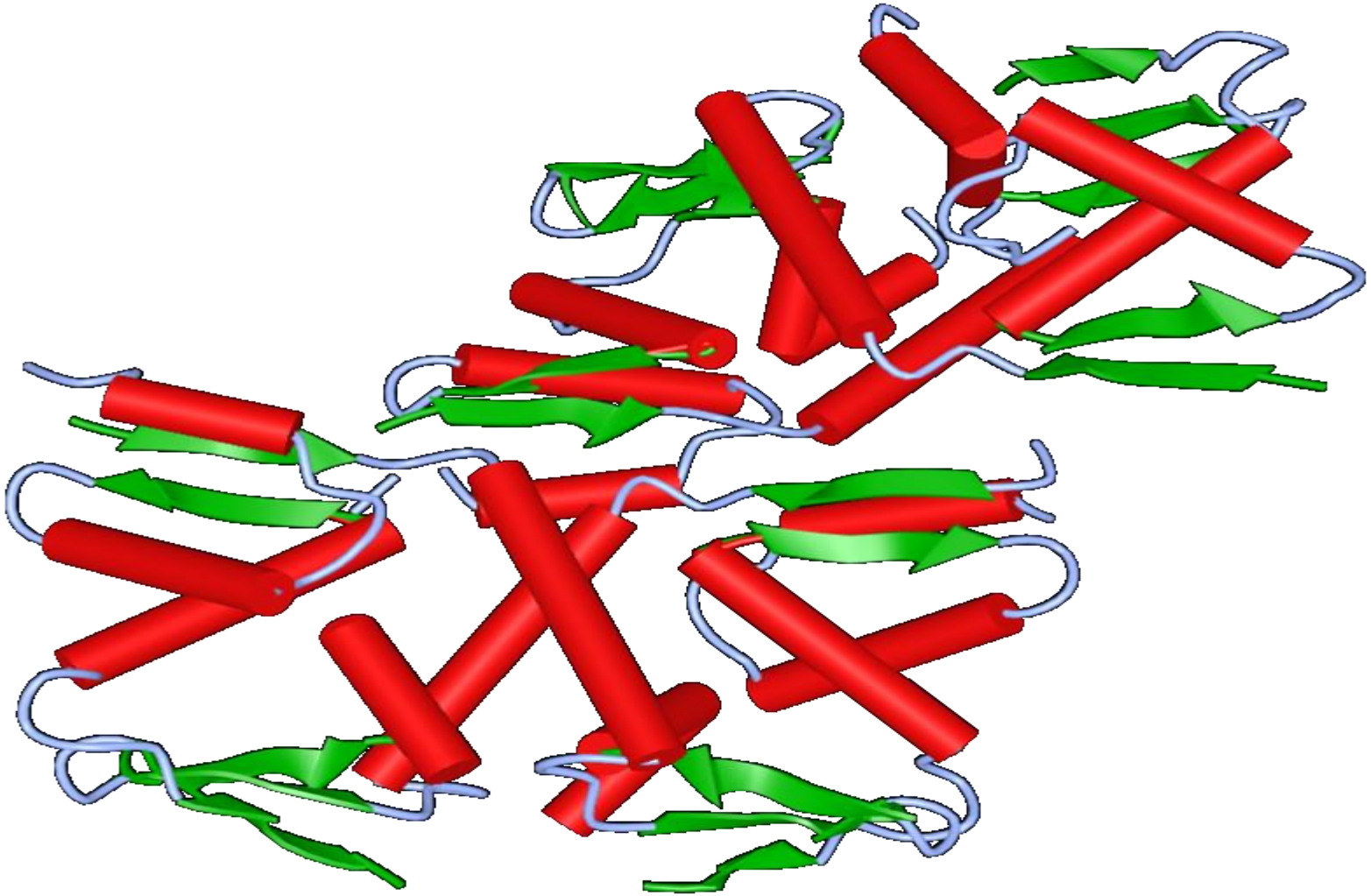
Now let us sum up the method as applied to the Bach protein: it has been identified in tetrapods, cartilaginous and bony fishes, cyclostomes, and ascidians using a local alignment with the human Bach1, the presence of the BTB and bZIP-Maf domains, satisfactory alignment of the secondary structure, and high similarity of the 3D structure at the N- and C-terminal regions with the human Bach1.



The possible absence of Bach in *Oikopleura dioica*, which is relatively close to *Ciona* spp., can be attributed to its neoteny. Indeed, like other appendicularians, adult *O. dioica* has a discrete body and tail and preserves the notochord throughout its life, while the body structure of *Ciona* spp. substantially changes in development.

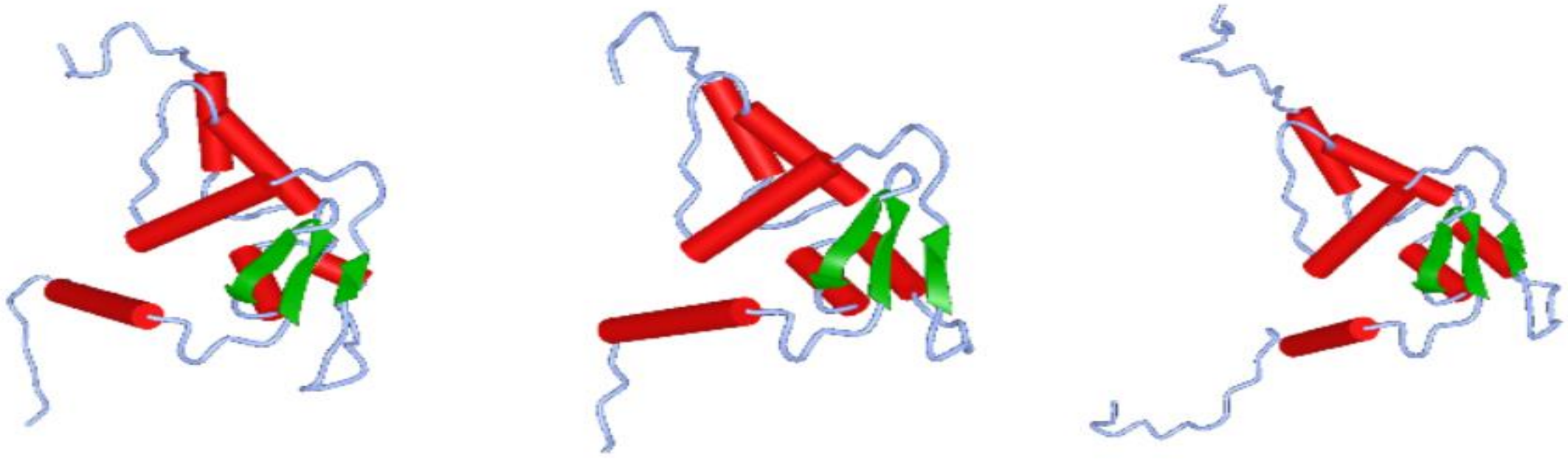
We recall **the tertiary structure of the BTB domain of human BACH1**

(4 chains):



The **tertiary** **structures of the BTB** domain in **Bach1** (single chain)

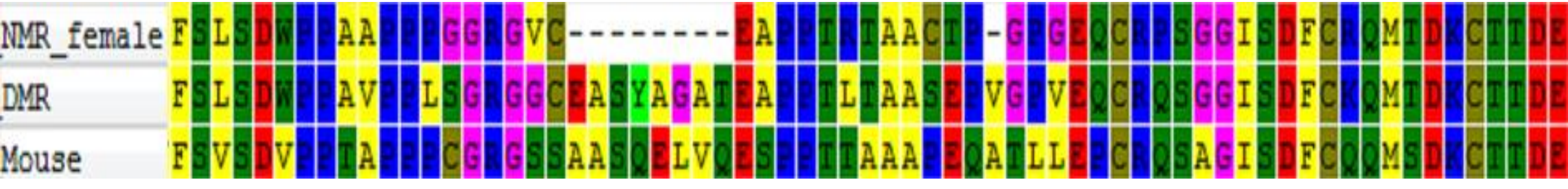
in ***Ciona intestinalis***, ***Ciona savignyi***, ***Eptatretus burgeri***



Within Euarchontoglires, Bach1 demonstrates only minor changes in most rodents including the Damaraland mole-rat (DMR) *Fukomys damarensis*. The **only exception is the C-terminal region of all Bach1 isoforms in the naked mole-rat** (NMR) *Heterocephalus glaber*.

One Bach1 isoform has a deletion, although a short region upstream of the C-terminus is conserved, **see here**. **Another isoform** has a very long insertion in the same region. The changes in the NMR Bach1 atypical for rodents can also be related to the unusually long lifespan.

The full-length Bach1 protein of the NMR usually shares (variable) amino acids with the DMR but not with the mouse, which agrees with their taxonomical position.



No significant differences in Bach1 have been revealed in primates.
Beyond Euarchontoglires, the Bach1 sequences of the sloth *Choloepus hoffmanni,* tenrec *Echinops telfairi*, shrew *Sorex araneus*, dolphin *Tursiops truncatus*, and flying fox *Pteropus vampyrus* are similar to those in primates and most rodents.

Conversely, other representatives of Laurasiatheria as well as the nine-banded armadillo *Dasypus novemcinctus* have insertions or deletions in the same region of Bach1 as in NMR *Heterocephalus glaber*.

The Bach1 of the hedgehog *Erinaceus europaeus*, elephant *Loxodonta africana*, and hyrax *Procavia capensis* has a very long C-terminal deletion covering both conserved and variable regions.

Let us continue comparing the **Bach1** and **Bach2** proteins.

In jawed vertebrates (Gnathostomata), the homodimer formation was found to involve interactions of the kinked N - terminus (**N - hook**) and the partner's **C - terminal residues in Bach1**:

```
mouse Bach1   1 MSVSE...SAVFA 10
human Bach1   1 MSLSE...NSVFA 10
mouse Bach2   1 MSVDEKPGSPVFA 13
human Bach2   1 MSVDEKPDSPMYV 13
```

The sequence logo of **N-termini in Bach2**:

The CP-motif **223-LCPKYR-228 inside the Bach1** protein contains substitution in marsupials as well as in some lizards (essential for heme binding):

Mus musculus	LCPKYR	
Monodelphis domestica	LGPKYR	Marsupials
Sarcophilus harrisii	LGPKYR	
Phascolarctos cinereus	LGPKYR	
Vombatus ursinus	LGPKYR	
Ornithorhynchus anatinus	LCPKYR	
Pogona vitticeps	QCPKYR	Lizards
Podarcis muralis	LYPKYR	
Lacerta agilis	LYPKYR	
Anolis carolinensis	QCPKYR	
Gekko japonicus	LCPKYR	
Thamnophis elegans	LCPKYR	
Thamnophis sirtalis	LCPKYR	
Notechis scutatus	LCPKYR	
Pseudonaja textilis	LCPKYR	
Python bivittatus	LCPKYR	

Thus, the predicted heme-binding motifs in Bach1 of tetrapods insignificantly differ from those in mouse **excluding the species specified** (≥5 CP-motifs):

223-LCPKYR-228 (C→G in marsupials, although it is the same in the mouse and platypus *Ornithorhynchus anatinus*; or C→Y in the common wall lizard *Podarcis muralis*);

300-QCPAEQ-305, which considerably changed or disappeared in most mammals;

435-ECPWLG-340 (conserved in all tetrapods);

463-NCPFIS-468 (the cysteine is conserved in tetrapods; and I→M, in placentals including DMR, NMR and the common degu *Octodon degus*);

494-PCPYAC-499 (conserved in all tetrapods except the bearded dragon *Pogona vitticeps* and platypus);

648-DCPLSF-653 (conserved in almost all tetrapods).

Now about **CP motifs for heme-dependent regulation**.

The **Bach** proteins **in *Ciona spp*.** have **ONLY two conserved CP motifs with each other. They DIFFER** from those involved in the **heme-dependent regulation** **in human and mouse** that corresponds early diverging of *Ciona*:

Ciona savignyi CKNSKGDCPLMAKLSL
Ciona intestinalis VKNSKGDCPLMAKLSL

Ciona savignyi CSVMSQACPMMSQACT
Ciona intestinalis CSVMSQSCPMMSSPCS

In human ***Homo sapiens*** and elephant shark *Callorhinchus milii* (**chimera**) **the Bach1 proteins have three conserved** (for these species) **CP motifs involved in the heme-dependent regulation.** But in hagfish *Eptatretus burgeri* (**Myxini**) there are only two such motifs common with humans**.** Thus, the quantity of heme-binding motifs decreases but stay sufficient (6 > 5 > 3 > 2).

```
Homo sapiens          ALALPSLCPKYRKFQKAFG
Callorhinchus milii   QSRWASLCPKYRKFQLACG
Eptatretus burgeri    SCTSPLKCPTTRSLCLETQ

Homo sapiens          TTLSSVNCPFISTLSTEGC
Callorhinchus milii   SSLSSSKCPFGYTAGSSVC
Eptatretus burgeri    HVGDSPSCPINLSLACKDP

Homo sapiens          AKYSAADCPLSFLISEKDK
Callorhinchus milii   AKYSSPECPLSVLNIQRST
Eptatretus burgeri    --------------------
```

The **Bach2** proteins have **CP motifs that are conserved in most vertebrates.** They holds for the West Indian Ocean coelacanth ***Latimeria chalumnae*** **(kindred of tetrapods)**, **cartilaginous fishes**: the elephant shark *Callorhinchus milii*, the whale shark *Rhincodon typus*, the thorny skate *Amblyraja radiata* (**jawed vertebrates**). In Latimeria unusual C→Y.

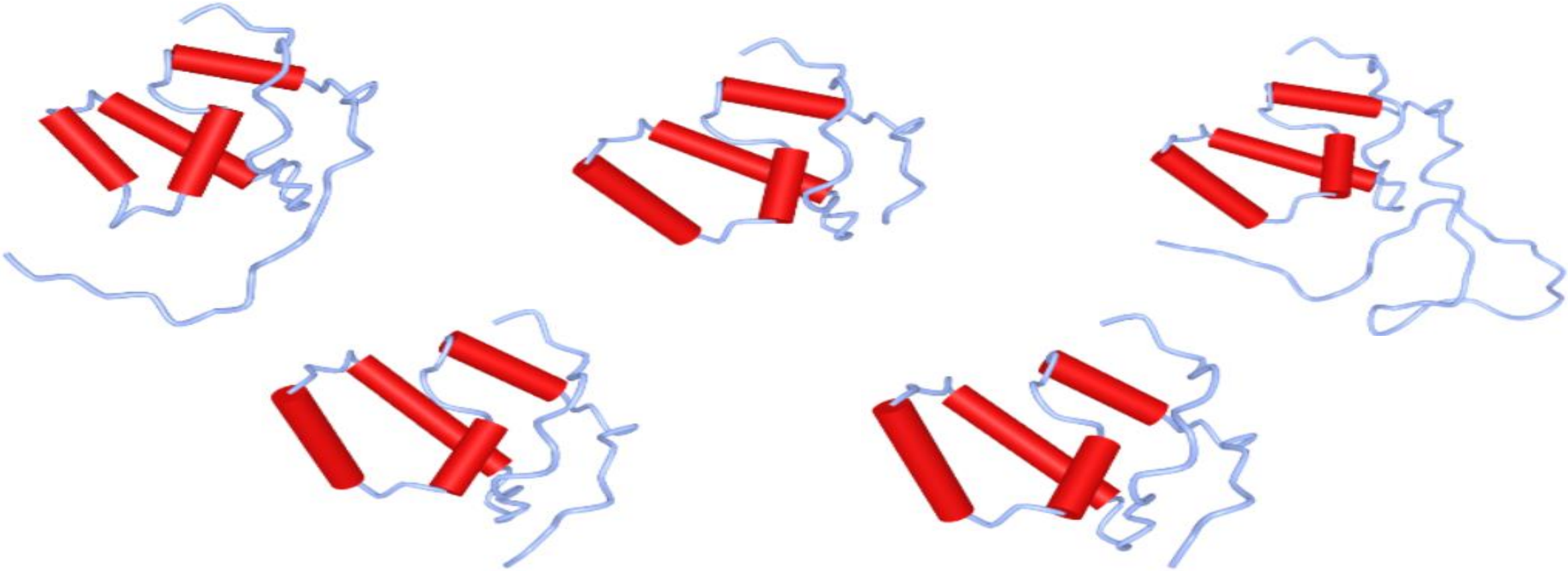
```
Homo sapiens          NTSCPVPIKVCPRSPPLETRTRTSSSCSSYSYAEDGSG ...
Mus musculus          NTSCPVPIKVCPRSPPLETRTRTSSSCSSYSYAEDGSG ...
Latimeria chalumnae   NTSCPVPIKVYPRSPPLETRTRTSSSCSSYSYAEDGSG ...
Callorhinchus milii   NTSCPVPIKVCPRSPP-ETRTRTSSSCSSYSFPEDGSG ...
Rhincodon typus       KTSCPVPIKVCPRSPPSETRTRTSSSCSSYSFPEDGSG ...
Amblyraja radiata     NTSCPVPIKVCPRSPPSETRTRTSSSCSSYSFPEDGSG ...

Homo sapiens          SPEQIQALHRYCPVLRPMDLPTASSINPAPL-GAEQNI ...
Mus musculus          SPEQIQALHRYCPVLIPMDLPGAS-VNPPPV-GVEQSL ...
Latimeria chalumnae   SPEQIQALHRYCPVLRPMEQPVTASIDPSLS-LLEQSL ...
Callorhinchus milii   SSEQIQSLHRYCPALRPLDQAATAGSDTSPSAEFEQKP ...
Rhincodon typus       NPEQIQSLHKFCPVFRPMDESAAKASTTLPLAGLEQKI ...
Amblyraja radiata     SPEQMQT---FCPVFRPVDEPAATGSTISPPDGLERKF ...
```

Now let us sum up the method as applied to Nrf2:

the **Nrf2** protein has been identified in deuterostomes **considering** a local alignment with the human NFE2L2 (NP_006155.2), the presence of the bZIP-Maf domain, a satisfactory alignment of the secondary structure, the absence of some “wrong” domains (kelch-type beta-propeller in kelch-like proteins, BTB in Bach, bZIP-Jun in Jun and similarly for other bZIP proteins),

and a high 3D structure similarity with NFE2L2 at the C-terminus coupled with a significant difference from other human bZIP proteins at the N-terminus.

As well as the **tertiary** structures of DNA-binding domain in **Nrf2** in ***Homo sapiens***, *Branchiostoma belcheri* (Chordata), *Saccoglossus kowalevskii* (Hemichordata), *Strongylocentrotus purpuratus* (Echinodermata), *Drosophila melanogaster* (Protostomia) are quite similar:



Thank You

The study was funded by the RFBR, project no. 18-29-13037