Scientific Research Publishing

Computational Molecular Bioscience, 2022, 12, 58-83
https://www.scirp.org/journal/cmb
ISSN Online: 2165-3453
ISSN Print: 2165-3445

Antibody-Like Phosphorylation Sites. Theme for Studies of Cancer, Aging and Evolution

Jaroslav Kubrycht[1*], Karel Sigler[2]

[1]Department of Physiology, Second Faculty of Medicine, Charles University, Prague, Czech Republic
[2]Laboratory of Cellular Biology, Institute of Microbiology, Academy of Sciences of the Czech Republic, Prague, Czech Republic
Email: *jkub@post.cz, sigler@biomed.cas.cz

How to cite this paper: Kubrycht, J. and Sigler, K. (2022) Antibody-Like Phosphorylation Sites. Theme for Studies of Cancer, Aging and Evolution. *Computational Molecular Bioscience*, **12**, 58-83.
https://doi.org/10.4236/cmb.2022.121004

Received: February 11, 2022
Accepted: March 28, 2022
Published: March 31, 2022

Abstract

Sequence similarities were found between protein and DNA sequences encoding certain part of conserved variable immunoglobulin domains (*i.e.* conserved IgV) and phosphorylation sites. Hypermutation motifs were then indicated in the majority of the corresponding non-IgV nucleotide sequences. According to database confirmations or double prediction of phosphorylation sites, 80% of the selected human and mouse IgV-related phosphorylation sites or their highly probable candidates exhibited substrate relationship to ataxia-telangiectasia-mutated kinase known as ATM. In accordance with literature data, inactivation of ATM by mutations can participate in the mechanisms of carcinogenesis, neurodegeneration and possibly also in aging. In agreement with this relationship, some of the selected IgV-/ATM-related segments formed molecules specifically involved in carcinogenesis. The selected IgV-related sequence segments were also similar to certain segments of higher plants containing immunoglobulin-like repeats and related regions. Bioinformatic analysis of some selected plant sequences then indicated the presence of catalytic domains composing serine/threonine/tyrosine receptor/receptor-like kinases, which are considered important structures for evolution of very early and part of later Ig-domain-related immunity. The analyzed conserved domain similarities also suggested certain interesting structural and phylogenic relationships, which need to be further investigated. This review in fact briefly summarizes the findings on the subject from the last twenty years.

Keywords

Ataxia-Telangiectasia-Mutated Kinase (ATM), Carcinogenesis, Complementarity Determining Region 1 (CDR1, Hypervariable Region 1), Conserved Domain(s), Deep Evolution, Evolution, Hypermutation, Kinase(s),

Phosphorylation Site(s), Plant Immunity, Variable Immunoglobulin Domain(s) (IgV)

1. Introduction

Segments of immunoglobulin variable domains (**IgV**) of immunoglobulins (**Ig**) or T-cell receptors interacting with antigens as well as protein phosphorylation sites represent short peptides whose interactions can be considerably altered by one or a few non-synonymous mutations in the corresponding encoding DNA. Therefore, the question was whether at least some of the phosphorylation sites were not structurally close to IgV segments [1] [2]. In the initial examination of this question, our chain matrix was used [3], and intertwined and linear similarities were found between 1) certain model protein kinase substrates or inhibitors and 2) different IgV consensi [1] [2] [4]. The N-terminal IgV region containing the C-terminal part of FR1, the **hypervariable CDR1** region and the N-terminal part of FR2 appeared to be the most structurally interesting in terms of the sought similarities and also frequent hypermutation [2] [4] [5], while the greatest similarities between non-vertebrate metazoan IgV and also between different conserved IgV were found in the FR3 framework (this fact could be rather important for the evolution of recombination of antibody genes [5] [6]). Nevertheless, not only the selected N-terminal regions but also such **FR3 segments** of different conserved IgV constructs exhibited dominant occurrence of predicted phosphorylation sites [7].

The corresponding more detailed search for **IgV-related murine and human phosphorylation sites** was then performed using 1) sequences of the conserved IgV domains corresponding to N-terminal region mentioned above, 2) bioinformatic procedures analyzing simultaneously the corresponding RNA and protein sequences (*i.e.* bilingual approach including mainly predictions based on artificial intellect and searches for hypermutation motifs) and 3) databases of existing phosphorylation sites [6] (for details see next chapter). This study and some related attempts mentioned above thus belonged to the medical studies of mutability or potential mutability of regulatory important phosphorylation sites. The number of such studies increased mainly in last ten years [8]-[13].

Existing and convincingly predicted phosphorylation sites also contributed to our attempt to search for Ig-domain-related structures in higher plants [14] prolonging our part of research intended to **evolution** [4] [5]. These Ig-domain-related aims were accompanied with our little contribution to long lasting research specifically concerning **deep evolution** of the catalytic domains of serine/threonine/tyrosine kinases [15] [16] [17] [18] [19] in the plant kinase molecules containing Ig-like segments [14]. As well known, some of these kinases often accompany IgV domains in non-vertebrate metazoan proteins (cf. [5] [14]). For more detailed comments to our phylogenic studies see the chapter 3.

Common features of the informatic processes used in our sequence-based investigations are described in Figure 1.

2. IgV-Like Phosphorylation Sites and Their Relatives Found in Human and Mouse Sequences

As was specifically described in the corresponding our paper [6] (cf. also Figure 1), several types of BLAST searches for IgV-related segments occurring within sequences different from antigen receptors were performed when using two multiple nucleotide sequence queries (**MNQ**) and score-derived or combined limits. In summary, these MNQ were formed by preselected 149 different conserved N-terminal non-redundant IgV segments of reference mRNA sequences encoding Ig and T-cell receptors. Following the searches with MNQ, anti-redundant procedures finally restricted the set of six hundred ten IgV-related nucleotide sequence segments.

To predict phosphorylation sites in the next selection step, the originally selected sequences were transformed to the corresponding protein sequences. The prediction was realized by means two methods based on artificial intellect [6].

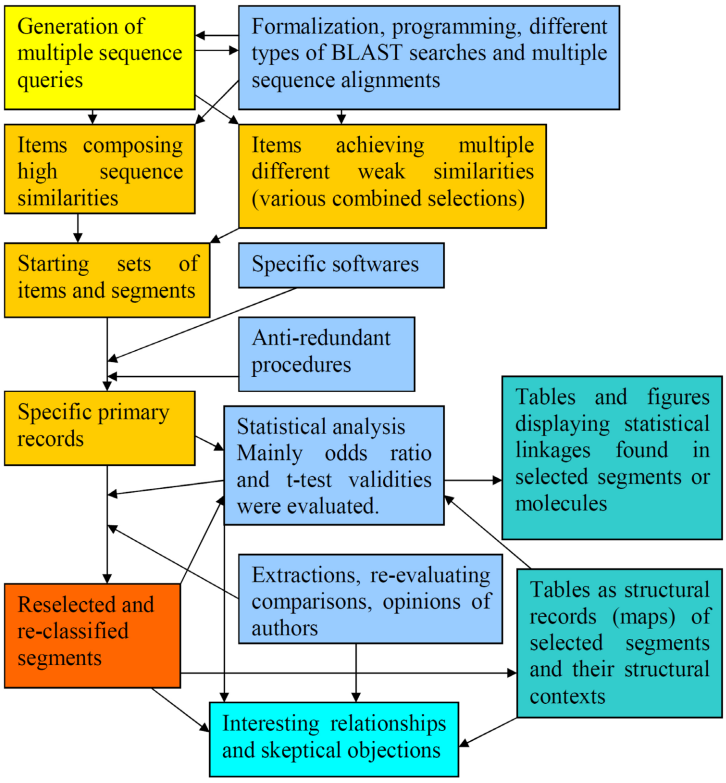

Figure 1. Common features of the selection and evaluation processes employed in our previous studies of human, mouse and plant sequences. For the selected associated results see the chapters 2 and 3 of this paper. For the specific schemes and more extended comments see the previous papers [6] [14]. For a more detailed description of the necessary new approaches and applied statistics see the corresponding paper supplements accessible on http://www.papersatellitesjk.com/.

More precisely, we used online programs KinasePhos2.0 and NetPhos2.0 working on the principles of support vector machines and neural networks, respectively [20] [21]. These programs usually selected most probable alternatives of substrate peptide segments for several kinases or phosphatases. This means that the programs mostly indicated fused chimerical regions of predicted phosporylations (**FCRP**) with close amino acids predicted as phosphorylated in the tested sequences. Such regions looked like **partially realized or only sleeping pluri-pontent germ-like structures** rather than sole phosphorylation sites. Primarily, all peptide segments achieving at least one prediction score 0.800 were restricted. Subsequently, we required presence of at least one certain tri- or tetranucleotide hypermutation motifs denoted here as HM* in the regions encoding superiorly predicted phosphorylation sites. These **HM*** were defined by 1) their location at positions critical with respect to possible aa alteration, *i.e.* as motifs associated with non-synonymous mutations and 2) the corresponding structure known from publications dealing with hypermutation of antigen receptors or hypermutation activities of oncologically important APOBEC enzymes (cf. [6] [22] [23]).

In the next step, we required either experimental confirmation published in databases (databases PHOSIDA and Phospho.ELM [24] [25] were applied in this case) or simultaneous achievement of the limiting score by the both predicting methods. Seven protein sequences forming the set of sequences denoted as **ALPS** (*i.e.* antibody-like phosphorylation sites sometimes interpreted as FCRP due to predictions) were restricted by means the databases. Thirty five protein sequences different from ALPS composed set of sequences called **ALPS2** (set of ALPS relatives mostly assembling FCRP) restricted only by successful double predictions. This set included among others two pairs of identical human and mouse sequences occurring in the corresponding orthologues. Fusion of ALPS and ALPS2 then yielded set of forty-two protein sequences entitled as **ALPS+** above all dealt with the corresponding paper [6].

The **products of the corresponding maximum score pairs** were individually enumerated for each ALPS+ by KinasePhos2.0 and NetPhos2.0. These products were distributed bimodally (formed two peaks), not as might be expected, with exponential decrease. More detailed statistical evaluation then indicated a strong and significant prevalence ($p < 0.05$) of these products to higher values than value 0.880 (limit for two identical scores 0.938) constituting the inferior border product value for dominant peak of product frequencies in range of maximum possible product values [6]. This result suggests the existence of selective evolutionary pressure that probably maintains the function of ALPS+ of the dominant peak and thus also attests to their validity of prediction. Consequently, ALPS and ALPS+ which product values formed the dominant peak composed set of **thirty seven** functionally correlated ALPS+ items (including also two identical sequence pairs described above in case of ALPS2). ALPS+ of this set were called here as **ALPS++**. IgV-related protein sequences including these ALPS++ were

used in the next phylogenic studies [14] (see also chapter 3). Some ALPS++ selected by authors are shown in **Table 1** (for associated sequencing projects see [26]-[41] and NCBI database of nucleotide sequences).

Since IgV-related segments including ALPS+-encoding sequences were primarily searched using comparison with the representative nucleotide sequences of MNQ, not all of the corresponding peptide candidates for ALPS+ were found in the same **reading frame** as the antigen receptor sequences being compared. In spite of it, considerable number of the ALPS+ amino acid sequences translated outside the reading frame of antigen receptors were identified as substrates of almost the same phosphorylating enzymes like ALPS+ translated in the same frame as IgV [6].

In our opinion, this remarkable independence of the reading frame could be of interest for understanding the evolution of **recognition plasticity** by IgV of antigen receptors compared to ALPS+ (cf. for instance MHC context in recognition by T-cell receptors). In addition, the observed differences in the reading frame appear to be important in terms of possible **epitope diversification** after HM*-mediated mutation. While translation of the mutated code for ALPS+ in a different reading frame from antigen receptors could lead to neo-epitopes, the translation in the same frame more likely causes cross-interactivities with rheumatoid autoantibodies [6].

Despite independent selection for kinase specificity, **80 percent** of the selected ALPS+ segments were predicted as substrates for ataxia-telangiectasia-mutated kinase (**ATM**) phosphorylating substrate on serine. Interestingly, inactivating mutations of the gene encoding ATM inhibit DNA damage response concerning first of all double strand breaks and lead to oxidative stress, genomic instability, disregulation of mitochondrial homeostasis and autophagy [42] [43] [44]. Such changes then cause cancer (mainly lymphomas), neurodegeneration, immunodeficiency, chronic lung disease and segmental premature aging [43] [45] [46] [47] [48]. Consistently with this fact, more than a quarter of the selected molecules containing ALPS+ were among those important in oncogenesis (cf. **Table 1**). Besides ATM-related ALPS+, we found among others ALPS+ of relationship to kinase **Aurora**, *i.e.* kinase that directly regulates the activity of the most frequently predicted ATM via its phosphorylation. In addition to ATM-phosphorylated **serine**, phosphorylations of **threonine** and tyrosine were also predicted in the evaluated IgV-related protein sequence segments. In summary, the results concerning predictions of phosphorylating enzymes were consistent with the previously predicted occurrence of phosphorylation sites in the corresponding compared N-terminal parts of **model conserved IgV** [7].

Several notable relationships followed from the statistical evaluation of **HM* occurrence in the corresponding DNA sequences** [6]. Only six human and four mouse DNA segments encoding ALPS+ included HM* in template (RNA complementary, *i.e.* directly transcribed) strands. In addition, the ratio between the occurrence of HM* and complementary group of hypermutation motifs was

Table 1. Selected existing or extremely probable ALPS++.

Title of molecule[a]	Sp[a]	Clonal names of sequences[b]	IgV-like segments	IgV-like similarities	aa_ HM*[b]	Prediction or database confirmation of phosphorylation sites			
		NS/PS	Position of NS/PS[b]	-max BS[c] -FrS[c]		pos_aa[a]	Score NetP[a]	KinP[a]	Kinase species[d]
Doublecortin-like kinase 1 [26] [27] [28] [29] [30]	Mu	XM_006500983.1 BAC41418.1	1343 - 1323 359 - 365	30.1 -	3	11S***PF** 11S 13S 16S	- 0.994 0.997- - 0.810- 0.879	- 0.970 0.942 0.924 0.982 PLK1	- GSK3 GSK3 ATM ATM
Tanslation factor BP (EIF4EBP1) [31]	Hu	AK312011.1 BAG34949.1	30 - 56 1 - 8	35.6 -	4	6S 8S***P**	0.944 -	0.982 -	ATM ATM
Limk2 (kinase) [26] [27] [28] [29] [30]	Mu	XM_006514552.1 XP_006514615.1	365 - 347 93 - 99	30.1 -	3	20S 20S***PF** 22S	0.991 - 0.979 -	0.907 - 0.945 0.891	ATM - Aur GSK3
Zinc finger protein 687 [32] [33]	Hu	NM_020832.1 NP_065883.1	1363 - 1388 422 - 430	35.6 -	3	23S***P** 23S	- 0.992 -	- 0.894 0.827	- ATM CK1
Ppp1r18/phostensin (phosphatase) [26] [27] [28] [29] [30]	Mu	XM_006536704.1 XP_006525157.1	907 - 889 241 - 247	30.1 -	2	13T***PF** 13T	- 0.976	- -	- -
Rps6ka4 (ribosomal PK) [26] [27] [28] [29] [30]	Mu	XM_006527231.1 XP_006527294.1	1573 - 1605 355 - 366	26.3* 7	2	16S***P** 16S	- - -	- 0.965 0.909	MAPK14 ATM MAPK
MAP3K1 (kinase) [34]	Hu	XM_005248520.2 Q13233.4	3795 - 3777 1401 - 1407	35.6 -	3	12T***PF** 9S	- - - - -	- 0.962 0.888 0.876 0.857	MAP3K1# Aur GSK3 CK1 ATM
Coiled-coil domain containing 88B [35]	Hu	XM_006718519.1 XP_006718582.1	1782 - 1808 578 - 587	35.6 -	2	14S 21S	0.996 - 0.990	0.986 0.877 0.925	ATM PLK1 ATM
Tight junction protein ZO-3 [26] [27] [28] [29] [30]	Mu	XM_006513708.1 XP_006513771.1	1214 - 1193 313 - 320	35.6 -	1	10S 14S 18S	0.998 0.998 0.997 -	0.954 0.984 0.977 0.888	ATM ATM ATM MAPK
Rps6kc1 (ribosomal kinase) [26] [27] [28] [29] [30] [36]	Mu	XM_006497177.1 NP_848890.3	671 - 701 214 - 224	30.1 -	4	11S 18S	0.995 0.951	0.975 0.959	ATM ATM
Mia3 (melanoma inhibitory) [37]	Mu	NM_177389.3 NP_796363.2	1771 - 1755 561 - 591	25.2* 16	2	14S 17S	0.958 - 0.994	0.957 0.881 0.970	ATM PLK1 ATM

Continued

Up-regulated in liver cancer 1 [38]	Hu	AB056749.1 NP_060177.2	2500 - 2538 819 - 831	28.1* 12	5	10S 24S	0.958 - 0.994	0.968 0.877 0.969	ATM PKB ATM
MAGEE1 (melanoma antigen) [39]	Hu	NM_020932.2 NP_065983.1	517 - 546 80 - 90	26.3* 6	4	20S	0.991 -	0.967 -	 ATM
Gltscr1 (tumor suppressor candidate) [40]	Mu	NM_001081418.1 NP_001074887.1	3610 - 3645 1168 - 1180	35.6 -	5	20S 22S	0.998 - 0.938 -	0.958 0.925 0.908 0.908	ATM GSK3 ATM GSK3
KDR = VEGF receptor 2 [41]	Hu	NM_002253.2 NP_002244.1	2343 - 2378 681 - 692	25.7* 42	5	9S 13S	0.975 0.984 -	0.977 0.921 0.849	ATM ATM PKB
Slowmo homolog 2 [26] [27] [28] [29] [30]	Mu	XM_006500002.1 XP_006500065.1	658 - 698 110 - 123	39.2 -	2	16S 19S	0.907 0.961 -	0.937 0.945 0.877	ATM ATM Aur

[a]The order of items was sorted according to 1) database confirmed existence and 2) products of score values obtained with ALPS++ predictions. **pos_aa**: predicted or database confirmed phosphorylated aa is denoted by a single character accompanied by the number indicating the corresponding aa position in the corresponding IgV-related peptide; ***F**, ***P**, ***PF**: database records confirming the existence of phosphorylation sites were found using Phosida, Phospho.ELM or both databases, respectively; **Hu**: human; **KinP**, **NetP**: scores obtained with prediction realized by KinasePhos2.0 and NetPhos2.0, respectively; **Mu**: mouse**; Sp**: species origin. [b]**aa_HM***: number of amino acids (**aa**), which compose existing or superior predicted phosphorylation site and are encoded by aa-altering nucleotide HM*; **NS/PS**: items in the two following rows concern nucleotide and protein sequences, respectively. [c]The similarities between pre-selected conserved-IgV-domain-related nucleotide sequences of vertebrate antigen receptors (**IgV-NS**) and human or mouse non-Ig sequences were searched. **Maximum local bit score** higher than thirty or maximum bit score higher than twenty five together with the presence of more than five supporting similarities were required. The records of supporting similarities had to comprise a) almost the same segment of subject sequence like the supported maximum similarity, b) different IgV-NS and c) score higher than 22 bits. For more detailed information see the corresponding original paper [6]. **FrS**: frequency of supporting similarities; **max BS**: maximum bit score of similarity with IgV-NS; **asterix**, *i.e.* ***after score values lower than 30**: effective frequency of supporting similarities was required. [d]**kinases specifically related to ALPS++ segments** #: autophosphorylation; **Aur**: Aurora-related kinase; **ATM**: ataxia-telangiectasia-mutated (kinase); **CK1**, **CK2**: casein kinases 1 and 2, respectively; **GSK3**: Glycogen synthetase kinase-3; **MAP3K**: MAPK kinase kinase; **PKB**, **PKC**: protein kinase B and C, respectively; **PLK1**: polo-like kinase 1; **dash**: confirmed phosphorylation was recorded by database though without the knowledge about the phosphorylating enzyme.

significantly and markedly higher in cases of **non-template (lagging) DNA** segments encoding ALPS+ (cf. Fig. 4c in [6]). These facts appeared to be interesting with respect to the observed predominant actions of mutator enzymes forming APOBEC3 family in extended non-template DNA of model bacteria and cancer cells [49]-[54]. Hence similar mutagenesis of oligonucleotide segments encoding actually phosphorylated ALPS+ could cause diminishing or loss of the corresponding phosphorylation or even exchange (cf. FCRP in chapter 2) of regulating phosphorylating enzyme. As well known such events sometimes lead to carcinogenesis [8] [9]. In addition, it is a question whether also the considered mutation changes of at least some frequent ATM-related ALPS+ can imitate consequences of ATM failure mentioned above and thus leading not only to

cancer but also to premature aging [6] [43] [45] [46] [47] [48]. In accordance with usual superior occurrence of HM* in segments of variable Ig genes encoding hypervariable regions 1 (CDR1), HM* occurred most frequently in ALPS+ encoded by nucleotide sequences similar to these CDR1. The twelve DNA segments encoding ALPS+ contained HM* whose specific change would directly remove **phosphorylated serine** from ALPS+.

3. Ig Domain Prehistory in the Focus of Conserved Domain Similarities

Higher plants represent *Bikonta* organisms. Consequently they differ from Unikonta group and its subgroup *Opsithokonta* including among others the taxonomic group of animals, *i.e. Metazoa* [55] [56] [57] [58], which frequently encode IgV in their genomes (cf. [5]). Hierarchically, **four types of similarities** were distinguished in our search for Ig-domain-related sequences of higher plants. The sets of protein sequences achieving the first type of similarity, *i.e.* **Ig-like similarity** (determining broad fraction of **Ig-like molecules**), were differently restricted from sixteen types of BLAST records including items of reference sequences. These starting BLAST records were prepared with the help of four sophisticated multiple-sequence queries (**MQ**) defined in our previous papers [6] [14]. More precisely, MQ were composed of two groups of specifically restricted highly conserved IgV segments (two **MQI**) and two groups of IgV-like protein segments including ALPS++ mentioned in previous chapter (two **MQP**). The selection of the discussed Ig-like similarities from initial set of BLAST records was performed by means of 1) collection of item samples achieving superior similarities (top10) in the limited short BLAST records or 2) sixteen different combined strategies from sixteen types of large BLAST records called mega-records [14]. Mega-records were only slightly limited by Expect values containing approximately 280 thousands of items. The additional three types of similarities then actually represented the next step in the selection of Ig-like molecules. In this step, 1323 pre-selected Ig-like molecules were evaluated using CDD software by comparing selected molecules with classified **conserved domains** (for CDD see papers [59] [60]). The two types of superior conserved domain similarities differed only in the existence of a lower Expect (E) limit. While **NRX similarities** were closed from below, $0.01 \leq E < 4.605$, **NRI similarities** were limited to only $E < 4.605$. The corresponding upper Expect limit 4.605 was compromised with respect to a) the large evolutionary distances of the higher plants and Metazoa, and b) the minimum value of Expect allowing a **significant rejection** of the evaluated conserved domain similarities and in fact also corresponding to dense conserved domain similarities between **fold length** segments [14]. Forty nine sequences achieving NRI similarities were found in our case. Nineteen of them were selected with help of MQP including sequences of ALPS++. **Significant conserved domain similarities** represented last type of the evaluated similarities. The Expect limit $E^* = 0.01$ for these similarities, cur-

rently applied in the selected widely-used CDD program [59] [60], was stricter than usual validity limit $p < 0.05$. Hence this limit restricted **significant probability** $p < w(E^*) = 0.0099502 \approx E^*$. In addition, independently verified **fold similarities** (fifteen sequences) and **contexts selected by the authors** (five sequences) also played a role in the final selection of the presented data (for overall list of the selected twenty items see Table 3 in the previous paper [14]).

Serine/threonine **protein kinase ALE2** of *Hevea brasiliensis* origin (sequence ID XP_021662681.1 [61]) containing a segment with significant conserved domain similarity to the vertebrate Ig1_Neogenin domain (**sIg1N** region at ALE2 positions 288-335) appeared to be the **most interesting molecule** in our study [14]. The identified conserved Ig1_Neogenin domain of this molecule usually indicates mammalian proteins involved in neurodegenerative processes, *i.e.* processes absent in plants. Since neurodegenerative processes in mammalians are associated with aging, we can ask whether this domain can also participate in aging processes in plants [14]. In addition to the observed significant similarity, there were four additional conserved Ig domain similarities at the NRX level and thirteen other such Ig similarities with Expect values in the interval $4.605 \leq E < 100$ (interval close to the range of dense similarities with the length of secondary structures) all co-locating with sIg1N. The simultaneously searched BLAST similarities of sIg1N with IgV-like segments including ALPS++ covered or overlapped most of sIg1N (ALE2 positions 292 - 303 and 305 - 335) were relatively weak (11.9 - 20.0 bits). Nevertheless, there were actually two repeating similarities with the same IgV-like segments including ALPS++ at ALE2 positions 305 - 328 and 329 - 359 (cf. web supplement of [14]). The corresponding repeat-associated ALPS++ composed ATM-related segment of a human protein annotated with sequence ID AB056749.1 ("Up-regulated in liver cancer 1" protein, *i.e.* URLC1 described by Okabe (2004) [38]). Similar co-localized occurrence of 1) NRI-similarities with conserved Ig domains and 2) similarities to IgV-like segments including ALPS++ were also found in the three cases of other molecules mentioned below (see **Table 2**).

Another significant conserved domain similarity comprising IG_FLMN (filamine Ig) domain was found in the gamete expressed 2 molecule (ID: XP_019237668.1). This similarity was overlapped by only slightly greater domain similarity with **filamine** and one NRX similarity with a **bacterial Ig domain**. The observed chimera of filamine and Ig domains appeared to be consistent with the data presented by Light (2012) [62] pointing to a possible common origin of filamines and immunoglobulins. In addition, significant conserved domain similarities of only bacterial Ig domains (simultaneously big_2 and BID_2) were found in the nuclear pore complex protein GP210 (ID: XP_010248630.1).

Among the molecules achieving **NRX-level of conserved domain similarity** with vertebrate conserved Ig domains and, at the same time, having required **fold similarities**, we have to mention first of all 1) molecule XP_010937019.2

Table 2. Some Ig-domain-related segments of higher plant origin.

Title/species/	Conserved domains[b]		Conserved domain similarities			AC-NRI[d]		HMMER as feedback[e]	
clonal name[a]	specification	access	Score	Expect	position[c]	num	position	Expect	WTG
nuclear	**Big_2**	pfam02368	50.86	**1.04E−07**	1154 - 1224	2	1149-	3.6E−13	*Bacteroidetes*
plex protein	**BID_2**	smart00635	41.23	**2.35E−04**	1149 - 1224	-	1224	6.9E−13	*Bacteroidetes*
Nelumbo nucifera	**BID_2**	smart00635	34.30	**0.06**	1561 - 1646	1	-	4.5E−16	*Bacteroidetes*
XP_010248630.1 [65]	**Big_2**	pfam02368	33.91	**0.08**	496 - 533	2	495 - 533	8.9E−17	*Proteobacteria*
Gamete expressed	Filamin	pfam00630	42.30	6.01E−05	92 - 222				
2, i. X3	**IG_FLMN**	smart00557	32.19	**0.20**	**92 - 228**	1	92 - 228	6.9E−13	*Archaea*
Nicotiana attenuate	**IG_FLMN**	smart00557	37.58	**2.89E−03**	**289 - 329**	3	252 - 329	**4.2E−27**	***Insecta***
XP_019237668.1 [66]	Filamin	pfam00630	37.68	2.61E−03	224 - 324				
**CoALIgdom	**BID_1**	smart00634	31.14	**0.45**	**252 - 316**	-	-	**1.8E−35**	***Proteobacteria***
STRK ALE2, i. X1	STKc_IRAK	cd14066	312.67	3.38E−98	765 - 1033				
Hevea brasiliensis	**Ig1_Neogenin**	cd05722	36.69	**9.26E−03**	**288 - 335**	5	288 - 335	**1.5E−27**	***Cnidaria***
XP_021662681.1 [67]	Atrophin-1 SF	cl26464	94.62	2.49E−19	33 - 490				
**CoALIgdom	PTKc_Src_like	cd05034	137.03	4.57E−36	763 - 961				
-	PTKc_VEGFR3	cd05102	65.39	5.05E−11	763 - 961				
LOC105056499	**Ig1_IL1R_like**	cd05756	32.44	**0.09**	**228 - 267**	1	-	8.8E−13	*Aves*
Elaeis guineensis									
XP_010937019.2									
[68] [69] [70]									
LOC25498589	DUF674	cl04913	404.72	3.55E−138	7 - 445				
Medicago truncatula	**IgV_H**	d04981	31.15	**0.45**	**299 - 353**	3	299 - 353	0.096	*Mollusca*
XP_013448976.1 [71]									
**CoALIgdom									
potassium channel	Ank_2 SF	cl26073	398.86	1.45E−127	65 - 697				
KOR2-like i.X1	**Ig1_PVR_like**	cd05718	28.93	**2.7**	**633 - 650**	2	604 - 655	6.8E−16	*Amphibia*
Musa acuminate	**ig**	pfam00047	28.31	**4.2**	**604 - 655**		-	**7.9E−51**	***Fungi***
XP_018679675.1 [72]									
LOC109149200	STKc_IRAK	cd14066	281.08	1.13E−85	505 - 770				

Continued

Ipomoea nil	**IGv**	smart00406	31.97	**0.40**	**53 - 106**	1	-	2.7E−03	*Cyanobacteria*
XP_019152409.1 [73]	B_lectin	cd00028	137.06	1.40E−37	29 - 151				
-	PTKc_Src_like	cd05034	141.27	2.18E−37	504 - 698				
-	PTKc_VEGFR3	cd05102	65.39	8.00E−11	1416 - 1589				
LOC108207541	Self-incomp_S1	pfam05938	104.22	5.96E−30	33 - 137				
Daucus carota	**Ig5_Contactin**	cd05852	26.52	**2.2**	**67 - 109**	1	-	1.7E−11	*Mollusca*
XP_017233469.1 [74] [75]	**-1**								
LRR STRK RCH1	PLN00113 SF	cl26793	497.45	7.31E−159	22 - 1075				
Hevea brasiliensis	**Ig3_L1-CAML**	cd05731	28.94	**2.9**	**429 - 457**	1	-	**8.9E−31**	***Fungi***
XP_021677865.1 [67]	PTKc_Src_like	cd05034	115.46	1.18E−28	798 - 1066				
**CoALIgdom	PTKc_VEGFR3	cd05102	68.08	7.04E−12	797 - 1081				
G-type lectin	STKc_IRAK	cd14066	282.62	1.77E−89	504 - 770				
S-STRK At2g19130	**IGv**	smart00406	28.12	**4.1**	**58 - 96**	1	-	1.6E−07	*Nitrospira*
Brachypodium	B_lectin	pfam01453	113.96	5.09E−30	76 - 193				
distachyon	PTKc_Src_like	cd05034	115.07	1.06E−28	503 - 696				
XP_010228837.1 [76]	PTKc_VEGFR3	cd05102	58.07	7.91E−09	504 - 694				

[a]For additional description of plant Ig-domain-related segments see Table 3 in [14]. **CoALIgdom: co-localized occurrence of 1) similarities with ALPS++ containing IgV-like segments (cf. Table 1) and 2) conserved Ig domain similarities. [b]Together with the monitored conserved Ig domain similarities (**Ig-cds**), we show here maximum conserved domain similarity present in each molecule-related CDD record and the molecular maxima related to similarities of catalytic kinase domains STYK-CD mentioned in the chapter 3. For more detailed information about the displayed conserved domains see the special option in the menu on NCBI web page. [c]**Pairs or triplets of restricted intervals present in grey elements** - **chimeras** of co-locating recessive conserved **Ig-cds** and dominant **non-Ig cds**. [d]**AC-NRI**: numbers and group-related maximum edge positions of all reciprocally co-locating or (if the number is one) solely found **Ig-cds**, whose Expects achieved at least NRI level (*i.e.* E < 4.605) mentioned in chapter 3. [e]We searched the maxima of **HMMER-derived similarities** between the selected Ig-domain-related segments of higher plant (*Embryophyta*) origin and the **complementary set** comprising sequences of all living organisms except for higher plants. **WTG**: well-known taxonomy group comprising species producing the most similar molecules restricted by HMMER searches; **bold**: Expect and taxonomy group of hot candidate segments for recent horizontal transfer (these segments were selected based on our empirical statistics; cf. the chapter 3 and [14]).

with a segment close to Ig-domains of IL1-receptors (unique molecule selected by three independent methods and in addition exhibiting significant specificity of top non-chimerical NRX similarity with the corresponding Ig domain), 2) an XP_013448976.1 molecule with maximum similarity to the IgV-H domain and two other colocalized Ig domain similarities at the NRX level, 3) a potassium channel molecule XP_018679675.1 selected by two NRX-limited Ig domain similarities. The similarities of potassium channel sequences to Ig domains have

so far been described in mammals [63] [64].

Furthermore, Kubrycht and Sigler (2020) [14] considered five molecules with an **interesting context**. Four of them formed **chimerical similarities** with vertebrate conserved Ig domains at the NRX level. This means that minor Ig-domain similarities were usually part of segments achieving more extensive dominant significant conserved domain similarities with non-Ig domains. This was the case for the dominant similarities with conserved domains of 1) B-lectins in molecules XP_019152409.1 and XP_010228837.1, 2) self-incomp_S1, *i.e.* domains causing so-called self-incompatibility preventing inbred pollination of plants (cf. XP_017233469.1 in **Table 2**) and 3) leucine-rich-repeat-associated kinases PLN00113 (cf. RCH1 molecule with ID: XP_021677865.1). In the last case, PLN00113-related segment separately contained in its two different parts two differently similar subsegments. More precisely, this concerned a) a more N-terminally located subsegment similar to the Ig3_L1-CAML domain, and b) a subsegment simultaneously significantly similar to many catalytic conserved domains of STY-kinases including such Ig-domain-associated catalytic domains (these catalytic kinase domains are otherwise described in the three terminal paragraphs of this chapter). For further details regarding the description of the Ig-domain similarities mentioned here see **Table 2** (for the sequencing projects associated with the displayed sequence items see [65]-[76] and NCBI database of protein sequences).

A **feedback database projection** of all Ig-domain related segments at the NRI similarity level was performed by Hidden Markov related HMMER and provided a bimodal frequency distribution within the scale of the logarithms of the corresponding Expect values (cf. **Table 2**). In agreement with this bimodal (two-peak) distribution the Expect limitation ($E < \mathbf{E^{\#}} = 10^{-26}$) was derived based on inferior values of the peak comprising the values determining superior similarities. Four Ig-domain related segments displayed in **Table 2** achieved the corresponding limited very high similarities (VHS). Due to unusually high similarity-related values of $E^{\#}$, these segments appeared to be hot candidates for products of recent **horizontal transfer** of genes or gene segments (cf. web supplement to [14] and Table 3 in the same paper). As can be expected based on simple skeptic objections assuming horizontal transfer of Ig domains from the metazoans, VHS to metazoan proteins included both selected peptide segments forming significant conserved domain similarities to vertebrate Ig domains, *i.e.* segments of ALE2 and Gamete expressed 2. More precisely, these important VHS were indicated when comparing the Ig-domain related segments of ALE2 and Gamete expressed 2 with proteins of stony corals *Pocillopora* (primitive metazoans belonging to clade *Cnidaria*) and insect origin, respectively [14]. On the other hand, two selected HMMER-derived VHS to sequences of *Fungi* origin can be classified evolutionarily interesting with respect to future supplementary investigation of Ig-domain-related molecules in yet belittled clades of *Opisthokonta* [77] [78]. These VHS were composed of already mentioned Ig-domain-related segments of the potassium channel and the RCH1 molecule (see the pre-

vious two paragraphs and Table 2).

An analysis of the sequences encoding multi-domain plant receptor/receptor-like kinases (**RKs**) provided also some orientation in the molecular evolution of signaling by cell-surface immune receptors [14]. These 121 sequences were restricted with the help of Ig-like similarities described above. Concerning the relationship to ALPS++, majority of RKs (104 of 121 sequences) were selected with combinatory contribution of IgV-related sequences including ALPS++, *i.e.* using always one of two MQP mentioned above. Consistently with records of significant conserved domain similarities or the literature, majority or at least some of the selected RKs appeared to be involved in plant antiviral immunity, respectively [14] [79]-[84]. In accordance with possible signaling role and name of RKs, their protein sequences frequently contained region or even regions achieving simultaneously many co-localized **robust** (extremely significant) conserved domain similarities with different model serine/threonine/tyrosine kinase catalytic domains (**STYK-CD**; cf. [15] [17] [19]). This means that the corresponding plant protein segments of RKs represented common similarity regions (**CSR**) which gave evidence about their undiversified structure with respect of indicated familial group of STYK-CD (a superfamily subset). Consequently, the conserved domain similarities indicated a **slowed down (or even frozen) evolution** of STYK-CD-related segments in plants.

The domain group of model STYK-CD included representative subset of nine special conserved tyrosine kinase domains (with robust average domain similarities in the range of 60 - 120 bits), which were often associated with IgV (**cdigvtk**) forming joint peptide chains of metazoan proteins [6]. In accordance with our data, cdigvtk represented primarily a reasonably defined domain set due to their model-validated IgV association in primitive *Metazoa* and minimized redundancy of the domain set (cf. section 2.9 and Fig. 4 in [14]). In 112 cases of the Ig-like RKs, these segments were significantly similar to at least one cdigvtk and in 102 cases to all nine cdigvtk. However, the other conserved domains than cdigvtk attained superior similarities with the STYK-CD-related segments of molecules from the set of RKs. This concerned mainly maximum similarities of CSR to conserved catalytic kinase domain associated with the interleukin 1 receptor (**IRAK**, *i.e.* cd14066; presented robust maximum mean bit score 286 bits) which occurred in 113 cases comprising CSR of 112 molecules participating in significant conserved domain similarities selected by cdigvtk. In accordance with the slowed down evolution of the STYK-CDs mentioned above, the dominantly similar IRAK could be assumed as a domain close to the molecular **ancestor of STYK-CD**, probably occurring earlier than the last common ancestor of plants and animals [14]. In addition, CSR of 118 RKs molecules, including CSR of 113 molecules significantly similar to IRAK, were significantly similar to the conserved domain of the leucine-rich-repeat-associated kinase PLN00113. The corresponding robust mean similarities with RKs were smaller than for IRAK but higher than for cdigvtk. However, certain segments immediately N-terminally adjacent to CSR of the compared plant sequences were sometimes also similar to

PLN00113. Such very robust extended similarities in some cases exceeded 450 bits, which corresponded to an overall similarity much higher than any IRAK domain similarity. This fact, together with the more frequent occurrence of the significant CSR-related conserved domain similarities in the set of RKs mentioned above, could indicate the similarity of PLN00113 with the **domain ancestor of STYK-CD even older than** the considered **ancestor close to IRAK** (cf. [14]). This would explain the lower conserved domain similarity of CSR to PLN00113 than to IRAK as a consequence of higher phylogenic distance when assuming scarcely maintained original functional importance of the critical N-terminally adjacent segments in some cases of PLN00113-related sequences.

The apparently controversial point in terms of the evolutionary relationship between the discussed IgV-associated cdigvtk domains and Ig domains was noticed. This point consisted in the fact that the set of forty nine molecules achieving superior Ig domain NRI similarities comprised only six or seven molecules with significant conserved domain similarities to all cdigvtk or at least one cdigvtk, respectively. The corresponding fraction was therefore significantly and markedly lower than the already mentioned fractions of significant cdigvtk similarities in the case of the set of RKs. Perhaps we could still consider the possibility of 1) slowed down development of STYK-CD and thus also their slower gradual evolution of gene-gene interactions with Ig domain ancestors or 2) interesting but not yet verified alternative of evolution from ancestor structures resembling some Ig-domain-related subsegment of PLN00113 to Ig domains (cf. RCH1 mentioned above and in **Table 2**).

4. Perspectives

The volume of sequence data and repertoire of the tools important for their processing keeps enlarging. In twenty years, super-smart computers can globally solve the problems we have listed here. Until then, however, we need to know a wide range of answers that will help physicians treat better and super-smart computers calculate, search or select better and more objectively.

We consider a more detailed study of the corresponding subset of **cancer-related molecules** to be a suitable for continuation of our effort. In this more detailed study, we would like to include the re-evaluation of selected segments using 1) database mapping of cancer-related mutations [85] [86] [87] [88] [89], tumor neo-epitopes [90]-[95] or other important relationships [96] [97] [98] and 2) prediction of cancer-driving sites [99] [100] or immunogenic epitopes [101]-[107] comprising also more specific cancer related neo-epitopes [108]. Based on literature data, the corresponding **phylogenic studies** could also include sequence sets of non-metazoan Unikonta (*Amoebozoa, Apusomonadida, Breviata*) more specifically evaluating non-metazoan *Opisthokonta* (including *Fungi* mentioned above). The reason for the proposed choice of the target organisms consists in the fact that a large number of early Ig domains and Ig-like segments occur in non-opisthokontal *Unikonta*, whereas structures close to IgV can be found in non-metazoan *Opisthokonta* [109] [110] [111] [112] (see also **Figure 2** and cf.

[113]-[120]). Due to our data and their statistical evaluation present in the preceding chapter and in [14], we assume that the corresponding future investigation could moreover expand to certain additional taxonomic groups of species than follows from literature. More precisely, this extension concerns yet not considered part of *Neozoa* representing close descendants of last common ancestor of higher plants and IgV-containing metazoans (cf. bottom branch following limiting dark brown division in the tree displayed in **Figure 2**). The

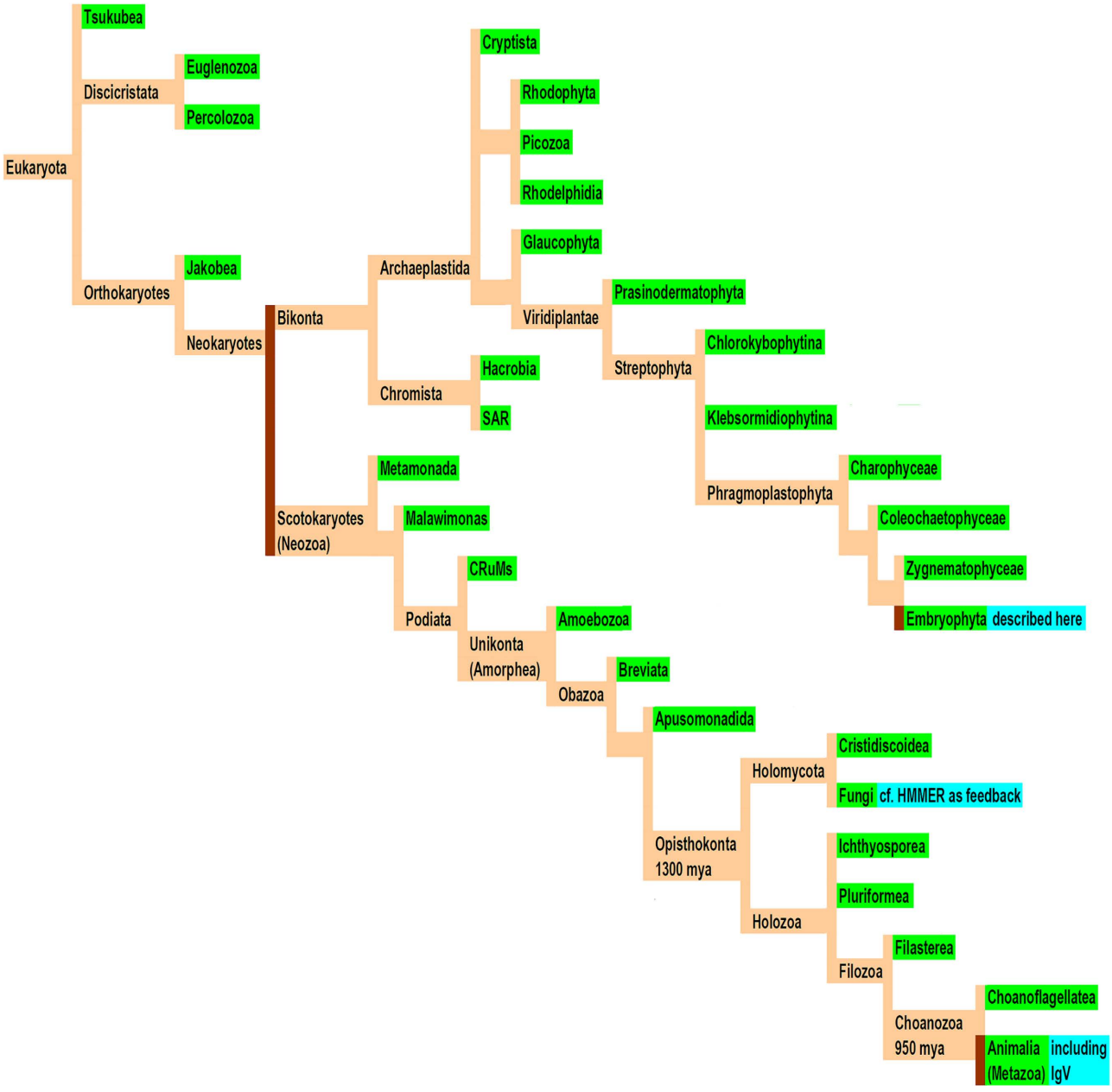

Figure 2. Phylogenic tree illustrating divergent evolution of investigated higher plants (*Embryophyta*) and metazoans expressing structurally important IgV (domains). This tree was constructed when comparing NCBI taxonomy [122] and the trees recently published in Wikipedia (see also the corresponding recent papers [113]-[120]). Blue: notes and comments; dark brown: branches or leaves substantial for the illustrated divergence; green: simplified restriction of terminal parts, *i.e.* leafs, of the displayed tree; mya: millions years ago.

proposed taxonomy-based extensions of compared sequences would require a feasible usage of more invasive and specific search strategies than those used in our previous paper [14], when respecting the recent frames of taxonomy relationships following from taxonomic databases [121] [122] and recent knowledge [58] [77] [123]-[130]. The possible relationships of at least some ALPS+ to **aging** have been also discussed here and in our last but one paper [6]. We suppose that also this context of our investigation will meet with better bioinformatic and more specific factual background in this decade.

We also hope that at least some of our methodological procedures and approaches, described so far mostly in publication supplements (e.g. our BLAST-related evaluation of multiple sequence alignments, an approximation determining the specificity of conserved domain similarities and several contributions to odds ratio evaluations), will be made publicly available in the future. In addition, the important methodic question for future consists evidently in deeper structural description of the observed sequence relationships.

Acknowledgements

Authors gratefully acknowledge their families.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Kubrycht, J. and Sigler, K. (1997) Animal Membrane Receptors and Adhesive Molecules. *Critical Reviews in Biotechnology*, **17**, 123-147. https://doi.org/10.3109/07388559709146610

[2] Kubrycht, J., Borecký, J. and Sigler, K. (2002) Sequence Similarities of Protein Kinase Peptide Substrates and Inhibitors: Comparison of Their Primary Structures with Immunoglobulin Repeats. *Folia Microbiologica*, **47**, 319-358. https://doi.org/10.1007/BF02818689

[3] Kubrycht, J. and Borecký, J. (1998) Matrix Formalization for Simple Sequence Comparison and Visualization of Short Sequence Relationships. *Imunologický Zpravodaj*, **27**, 21-27.

[4] Kubrycht, J., Borecký, J., Souček, P. and Ježek, P. (2004) Sequence Similarities of Protein Kinase Substrates and Inhibitors with Immunoglobulins and Model Immunoglobulin Homologue: Cell Adhesion Molecule from the Living Fossil Sponge *Geodia cydonium*. Mapping of Coherent Database Similarities and Implications for Evolution of CDR1 and Hypermutation. *Folia Microbiologica*, **49**, 219-246. https://doi.org/10.1007/BF02931038

[5] Kubrycht, J., Sigler, K., Růžička, M., Souček, P., Borecký, J. and Ježek, P. (2006) Ancient Phylogenetic Beginnings of Immunoglobulin Hypermutation. *Journal of Molecular Evolution*, **63**, 691-706. https://doi.org/10.1007/s00239-006-0051-9

[6] Kubrycht, J., Sigler, K., Souček, P. and Hudeček, J. (2016) Antibody-like Phosphorylation Sites in Focus of Statistically Based Bilingual Approach. *Computational Molecular Bioscience*, **6**, 1-22. https://doi.org/10.4236/cmb.2016.61001

[7] Kubrycht, J., Sigler, K., Souček, P. and Hudeček, J. (2013) Structures Composing

Protein Domains. *Biochimie*, **95**, 1511-1524.
https://doi.org/10.1016/j.biochi.2013.04.001

[8] Tram, E., Savas, S. and Ozcelik, H. (2013) Missense Variants of Uncertain Significance (VUS) Altering the Phosphorylation Patterns of BRCA1 and BRCA2. *PLoS One*, **8**, e62468. https://doi.org/10.1371/journal.pone.0062468

[9] Wang, Y., Cheng, H., Pan, Z., Ren, J., Liu, Z. and Xue, Y. (2015) Reconfiguring Phosphorylation Signaling by Genetic Polymorphisms Affects Cancer Susceptibility. *Journal of Molecular and Cellular Biology*, **7**, 187-202.
https://doi.org/10.1093/jmcb/mjv013

[10] Ma, S., Menon, R., Poulos, R.C. and Wong, J.W.H. (2017) Proteogenomic Analysis Prioritises Functional Single Nucleotide Variants in Cancer Samples. *Oncotarget*, **8**, 95841-95852. https://doi.org/10.18632/oncotarget.21339

[11] Lin, H.P., Ho, H.M., Chang, C.W., Yeh, S.D., Su, Y.W., Tan, T.H. and Lin, W.J. (2019) DUSP22 Suppresses Prostate Cancer Proliferation by Targeting the EGFR-AR Axis. *The FASEB Journal*, **33**, 14653-14667.
https://doi.org/10.1096/fj.201802558RR

[12] Liu, G., Weiner, H.L., Pederson, W.C., Davies, L. and Buchanan, E.P. (2019) Beta-Catenin Mutation with Complex Chromosomal Changes in Desmoid Tumor of the Scalp: A Case Report. *Craniomaxillofacial Trauma & Reconstruction*, **12**, 146-149.
https://doi.org/10.1055/s-0038-1676078

[13] Lin, S., Wang, C., Zhou, J., Shi, Y., Ruan, C., Tu, Y., *et al.* (2021) EPSD: A Well-Annotated Data Resource of Protein Phosphorylation Sites in Eukaryotes. *Briefings in Bioinformatics*, **22**, 298-307. https://doi.org/10.1093/bib/bbz169

[14] Kubrycht, J. and Sigler, K. (2020) Conserved Immunoglobulin Domain Similarities of Higher Plant Proteins. *Computational Molecular Bioscience*, **10**, 12-44.
https://doi.org/10.4236/cmb.2020. 101002

[15] Hanks, S.K., Quinn, A.M. and Hunter, T. (1988) The Protein Kinase Family: Conserved Features and Deduced Phylogeny of the Catalytic Domains. *Science*, **241**, 42-52. https://doi.org/10.1126/science.3291115

[16] Hirayama, T. and Oka, A. (1992) Novel Protein Kinase of *Arabidopsis thaliana* (APK1) that Phosphorylates Tyrosine, Serine and Threonine. *Plant Molecular Biology*, **20**, 653-662. https://doi.org/10.1007/BF00046450

[17] Rudrabhatla, P., Reddy, M.M. and Rajasekharan, R. (2006) Genome-wide Analysis and Experimentation of Plant Serine/Threonine/Tyrosine-Specific Protein Kinases. *Plant Molecular Biology*, **60**, 293-319. https://doi.org/10.1007/s11103-005-4109-7

[18] Klaus-Heisen, D., Nurisso, A., Pietraszewska-Bogiel, A., Mbengue, M., Camut, S., Timmers, T., *et al.* (2011) Structure-Function Similarities between a Plant Receptor-Like Kinase and the Human Interleukin-1 Receptor-Associated Kinase-4. *Journal of Biological Chemistry*, **286**, 11202-11210.
https://doi.org/10.1074/jbc.M110.186171

[19] Stancik, I.A., Šestak, M.S., Ji, B., Axelson-Fisk, M., Franjevic, D., Jers, C., *et al.* (2018) Serine/Threonine Protein Kinases from *Bacteria*, *Archaea* and *Eukarya* Share a Common Evolutionary Origin Deeply Rooted in the Tree of Life. *Journal of Molecular Biology*, **430**, 27-32. https://doi.org/10.1016/j.jmb.2017.11.004

[20] Blom, N., Gammeltoft, S. and Brunak, S. (1999) Sequence- and Structure-Based Prediction of Eukaryotic Protein Phosphorylation Sites. *Journal of Molecular Biology*, **294**, 1351-1362. https://doi.org/10.1006/jmbi.1999.3310

[21] Wong, Y.H., Lee, T.Y., Liang, H.K., Huang, C.M., Wang, T.Y., Yang, Y.H., *et al.*

(2007) KinasePhos 2.0: A Web Server for Identifying Protein Kinase-Specific Phosphorylation Sites Based on Sequences and Coupling Patterns. *Nucleic Acids Research*, **35**, W588-W594. https://doi.org/10.1093/nar/gkm322

[22] Kubrycht, J. and Sigler K. (2008) Length of the Hypermutation Motif DGYW/WRCH in the Focus of Statistical Limits. Implications for a Double-Motif or Extended Motif Recognition Models. *Journal of Theoretical Biology*, **255**, 8-15. https://doi.org/10.1016/j.jtbi.2008.07.039

[23] Roberts, S.A. and Gordenin, D.A. (2014) Hypermutation in Human Cancer Genomes: Footprints and Mechanisms. *Nature Reviews on Cancer*, **14**, 786-800. https://doi.org/10.1038/nrc3816

[24] Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J. and Diella, F. (2011) Phospho.ELM: A Database of Phosphorylation Sites—Update 2011. *Nucleic Acids Research*, **39**, D261-D267. https://doi.org/10.1093/nar/gkq1104

[25] Gnad, F., Gunawardena, J. and Mann, M. (2011) PHOSIDA 2011: The Posttranslational Modification Database. *Nucleic Acids Research*, **39**, D253-D260. https://doi.org/10.1093/nar/gkq1159

[26] Okazaki, N., Kikuno, R., Ohara, R., Inamoto, S., Hara, Y., Nagase, T., *et al.* (2002) Prediction of the Coding Sequences of Mouse Homologues of KIAA Gene: I. The Complete Nucleotide Sequences of 100 Mouse KIAA-homologous cDNAs Identified by Screening of Terminal Sequences of cDNA Clones Randomly Sampled from Size-Fractionated Libraries. *DNA Research*, **9**, 179-188. https://doi.org/10.1093/dnares/9.5.179

[27] Bayona-Bafaluy, M.P., Acín-Pérez, R., Mullikin, J.C., Park, J.S., Moreno-Loshuertos, R., Hu, P., *et al.* (2003) Revisiting the Mouse Mitochondrial DNA Sequence. *Nucleic Acids Research*, **31**, 5349-5355. https://doi.org/10.1093/nar/gkg739

[28] Church, D.M., Goodstadt, L., Hillier, L.W., Zody, M.C., Goldstein, S., She, X., *et al.* (2009) Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. *PLoS Biology*, **7**, e1000112. https://doi.org/10.1371/journal.pbio.1000112

[29] Steward, C.A., Humphray, S., Plumb, B., Jones, M.C., Quail, M.A., Rice, S., *et al.* (2010) Genome-Wide End-Sequenced BAC Resources for the NOD/MrkTac and NOD/ShiLtJ Mouse Genomes. *Genomics*, **95**, 105-110. https://doi.org/10.1016/j.ygeno.2009.10.004

[30] Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., *et al.* (2011) Modernizing Reference Genome Assemblies. *PLoS Biology*, **9**, e1001091. https://doi.org/10.1371/journal.pbio.1001091

[31] Maruyama, Y., Wakamatsu, A., Kawamura, Y., Kimura, K., Yamamoto, J., Nishikawa, T., *et al.* (2009) Human Gene and Protein Database (HGPD): A Novel Database Presenting a Large Quantity of Experiment-Based Results in Human Proteomics. *Nucleic Acids Research*, **37**, D762-D766. https://doi.org/10.1093/nar/gkn872

[32] Beausoleil, S.A., Jedrychowski, M., Schwartz, D., Elias, J.E., Villén, J., Li, J., *et al.* (2004) Large-Scale Characterization of HeLa Cell Nuclear Phosphoproteins. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12130-12135. https://doi.org/10.1073/pnas.0404720101

[33] Denoeud, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., *et al.* (2007) Prominent Use of Distal 5' Transcription Start Sites and Discovery of a Large Number of Additional Exons in ENCODE Regions. *Genome Research*, **17**, 746-759. https://doi.org/10.1101/gr.5660607

[34] Schmutz, J., Martin, J., Terry, A., Couronne, O., Grimwood, J., Lowry, S., *et al.* (2004) The DNA Sequence and Comparative Analysis of Human Chromosome 5.

Nature, **431**, 268-274. https://doi.org/10.1038/nature02919

[35] Kennedy, J.M., Fodil, N., Torre, S., Bongfen, S.E., Olivier, J.F., Leung, V., *et al.* (2014) CCDC88B is a Novel Regulator of Maturation and Effector Functions of T Cells during Pathological Inflammation. *Journal of Experimental Medicine*, **211**, 2519-2535. https://doi.org/10.1084/jem.20140455

[36] Distler, U., Schmeisser, M.J., Pelosi, A., Reim, D., Kuharev, J., Weiczner, R., *et al.* (2014) In-Depth Protein Profiling of the Postsynaptic Density from Mouse Hippocampus Using Data-Independent Acquisition Proteomics. *Proteomics*, **14**, 2607-2613. https://doi.org/10.1002/pmic.201300520

[37] Bosserhoff, A.K., Moser, M. and Buettner, R. (2004) Characterization and Expression Pattern of the Novel MIA Homolog TANGO. *Gene Expression Patterns*, **4**, 473-479. https://doi.org/10.1016/j.modgep.2003.12.002

[38] Okabe, H., Furukawa, Y., Kato, T., Hasegawa, S., Yamaoka, Y. and Nakamura, Y. (2004) Isolation of Development and Differentiation Enhancing Factor-like 1 (DDEFL1) as a Drug Target for Hepatocellular Carcinomas. *International Journal of Oncology*, **24**, 43-48.

[39] Albrecht, D.E. and Froehner, S.C. (2004) DAMAGE, a Novel Alpha-Dystrobrevin-Associated MAGE Protein in Dystrophin Complexes. *Journal of Biological Chemistry*, **279**, 7014-7023. https://doi.org/10.1074/jbc.M312205200

[40] Chae, T.H., Allen, K.M., Davisson, M.T., Sweet, H.O. and Walsh, C.A. (2002) Mapping of the Mouse *hyh* Gene to a YAC/BAC Contig on Proximal Chromosome 7. *Mammalian Genome*, **13**, 239-244. https://doi.org/10.1007/s00335-001-2144-5

[41] Terman, B.I., Carrion, M.E., Kovacs, E., Rasmussen, B.A., Eddy, R.L. and Shows, T.B. (1991) Identification of a New Endothelial Cell Growth Factor Receptor Tyrosine Kinase. *Oncogene*, **6**, 1677-1683.

[42] Stagni, V., Cirotti, C. and Barilà, D. (2018) Ataxia-Telangiectasia Mutated Kinase in the Control of Oxidative Stress, Mitochondria, and Autophagy in Cancer: A Maestro with a Large Orchestra. *Frontiers in Oncology*, **8**, Article No. 73. https://doi.org/10.3389/fonc.2018.00073

[43] Shiloh, Y. (2020) The Cerebellar Degeneration in Ataxia-Telangiectasia: A Case for Genome Instability. *DNA Repair*, **95**, Article ID: 102950. https://doi.org/10.1016/j.dnarep.2020.102950

[44] Stagni, V., Ferri, A., Cirotti, C. and Barilà, D. (2021) ATM Kinase-Dependent Regulation of Autophagy: A Key Player in Senescence? *Frontiers in Cellular and Developmental Biology*, **8**, Article ID: 599048. https://doi.org/10.3389/fcell.2020.599048

[45] Barzilai, A., Rotman, G. and Shiloh, Y. (2002) ATM Deficiency and Oxidative Stress: A New Dimension of Defective Response to DNA Damage. *DNA Repair*, **1**, 3-25. https://doi.org/10.1016/s1568-7864(01)00007-6

[46] Crawford, T.O., Skolasky, R.L., Fernandez, R., Rosquist, K.J. and Lederman, H.M. (2006) Survival Probability in Ataxia Telangiectasia. *Archives of Disease in Childhood*, **91**, 610-611. https://doi.org/10.1136/adc.2006.094268

[47] Boohaker, R.J. and Xu, B. (2014) The Versatile Functions of ATM Kinase. *Biomedical Journal*, **37**, 3-9. https://doi.org/10.4103/2319-4170.125655

[48] Choy, K.R. and Watters, D.J. (2018) Neurodegeneration in Ataxia-Telangiectasia: Multiple Roles of ATM Kinase in Cellular Homeostasis. *Developmental Dynamics*, **247**, 33-46. https://doi.org/10.1002/dvdy.24522

[49] Bhagwat, A.S., Hao, W., Townes, J.P., Lee, H., Tang, H. and Foster, P.L. (2016) Strand-Biased Cytosine Deamination at the Replication Fork Causes Cytosine to

Thymine Mutations in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 2176-2181. https://doi.org/10.1073/pnas.1522325113

[50] Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E. and Hess, J.M., *et al.* (2016) Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell*, **164**, 538-549. https://doi.org/10.1016/j.cell.2015.12.050

[51] Hoopes, J.I., Cortez, L.M., Mertz, T.M., Malc, E.P., Mieczkowski, P.A. and Roberts S.A. (2016) APOBEC3A and APOBEC3B Deaminate the Lagging Strand Template During DNA Replication. *Cell Reports*, **14**, 1273-1282. https://doi.org/10.1016/j.celrep.2016.01.021

[52] Morganella, S., Alexandrov, L.B., Glodzik, D., Zou, X., Davies, H., Staaf, J., *et al.* (2016) The Topography of Mutational Processes in Breast Cancer Genomes. *Nature Communications*, **7**, Article ID: 11383. https://doi.org/10.1038/ncomms11383

[53] Seplyarskiy, V.B., Soldatov, R.A., Popadin, K.Y., Antonarakis, S.E., Bazykin, G.A. and Nikolaev, S.I. (2016) APOBEC-Induced Mutations in Human Cancers Are Strongly Enriched on the Lagging DNA Strand during Replication. *Genome Research*, **26**, 174-182. https://doi.org/10.1101/gr.197046.115.

[54] Saini, N. and Gordenin, D.A. (2020) Hypermutation in Single-Stranded DNA. *DNA Repair*, **91-92**, Article ID: 102868. https://doi.org/10.1016/j.dnarep.2020.102868

[55] Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., *et al.* (2015) Bacterial Proteins Pinpoint a Single Eukaryotic Root. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E693-E969. https://doi.org/10.1073/pnas.1420657112.

[56] Plattner, H. and Verkhratsky, A. (2015) The Ancient Roots of Calcium Signalling Evolutionary Tree. *Cell Calcium*, **57**, 123-132. https://doi.org/10.1016/j.ceca.2014.12.004

[57] Beckmann, L., Edel, K.H., Batistič, O. and Kudla, J. (2016) A Calcium Sensor-Protein Kinase Signaling Module Diversified in Plants and is Retained in All Lineages of *Bikonta* Species. *Science Reports*, **6**, Article No. 31645. https://doi.org/10.1038/srep31645

[58] Plattner, H. (2018) Evolutionary Cell Biology of Proteins from Protists to Humans and Plants. *Journal of Eukaryotic Microbiology*, **65**, 255-289. https://doi.org/10.1111/jeu.12449

[59] Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y. and Bryant, S.H. (2002) CDD: A Database of Conserved Domain Alignments with Links to Domain Three-Dimensional Structure. *Nucleic Acids Research*, **30**, 281-283. https://doi.org/10.1093/nar/30.1.281

[60] Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., *et al.* (2015) CDD: NCBI's Conserved Domain Database. *Nucleic Acids Research*, **43**, D222-D226. https://doi.org/10.1093/nar/gku1221

[61] Li, D., Deng, Z., Qin, B., Liu, X. and Men, Z. (2012) *De Novo* Assembly and Characterization of Bark Transcriptome Using Illumina Sequencing and Development of EST-SSR Markers in Rubber Tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genomics*, **13**, Article No. 192. https://doi.org/10.1186/1471-2164-13-192

[62] Light, S., Sagit, R., Ithychanda, S.S., Qin, J. and Elofsson, A. (2012) The Evolution of Filamin—A Protein Domain Repeat Perspective. *Journal of Structural Biology*, **179**, 289-298. https://doi.org/10.1016/j.jsb.2012.02.010

[63] Fallen, K., Banerjee, S., Sheehan, J., Addison, D., Lewis, L.M., Meiler, J. and Denton,

J.S. (2009) The Kir Channel Immunoglobulin Domain is Essential for Kir1.1 (ROMK) Thermodynamic Stability, Trafficking and Gating. *Channels*, **3**, 57-68. https://doi.org/10.4161/chan.3.1.7817

[64] Yereddi, N.R., Cusdin, F.S., Namadurai, S., Packman, L.C., Monie, T.P., Slavny, P., *et al.* (2013) The Immunoglobulin Domain of the Sodium Channel β3 Subunit Contains a Surface-Localized Disulfide Bond that Is Required for Homophilic Binding. *The FASEB Journal*, **27**, 568-580. https://doi.org/10.1096/fj.12-209445

[65] Ming, R., Van Buren, R., Liu, Y., Yang, M., Han, Y., Li, L.T., *et al.* (2013) Genome of the Long-Living Sacred Lotus (*Nelumbo nucifera* Gaertn.). *Genome Biology*, **14**, Article No. R41. https://doi.org/10.1186/gb-2013-14-5-r41

[66] Bohlmann, J., Stauber, E.J., Krock, B., Oldham, N.J., Gershenzon, J. and Baldwin, I.T. (2002) Gene Expression of 5-*epi*-Aristolochene Synthase and Formation of Capsidiol in Roots of *Nicotiana attenuata* and *N. sylvestris*. *Phytochemistry*, **60**, 109-116. https://doi.org/10.1016/s0031-9422(02)00080-8

[67] Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., *et al.* (2016) The Rubber Tree Genome Reveals New Insights into Rubber Production and Species Adaptation. *Nature Plants*, **2**, Article No. 16073. https://doi.org/10.1038/nplants.2016.73

[68] Low, E.T., Alias, H., Boon, S.H., Shariff, E.M., Tan, C.Y., Ooi, L.C., *et al.* (2008) Oil Palm (*Elaeis guineensis* Jacq.) Tissue Culture ESTs: Identifying Genes Associated with Callogenesis and Embryogenesis. *BMC Plant Biology*, **8**, Article No. 62. https://doi.org/10.1186/1471-2229-8-62

[69] Uthaipaisanwong, P., Chanprasert, J., Shearman, J.R., Sangsrakru, D., Yoocha, T., Jomchai, N., *et al.* (2012) Characterization of the Chloroplast Genome Sequence of Oil Palm (*Elaeis guineensis* Jacq.). *Gene*, **500**, 172-180. https://doi.org/10.1016/j.gene.2012.03.061

[70] Singh, R., Ong-Abdullah, M., Low, E.T., Manaf, M.A., Rosli, R., Nookiah, R., *et al.* (2013) Oil Palm Genome Sequence Reveals Divergence of Interfertile Species in Old and New Worlds. *Nature*, **500**, 335-339. https://doi.org/10.1038/nature12309

[71] Pecrix, Y., Staton, S.E., Sallet, E., Lelandais-Brière, C., Moreau, S., Carrère, S., *et al.* (2018) Whole-Genome Landscape of *Medicago truncatula* Symbiotic Genes. *Nature Plants*, **4**, 1017-1025. https://doi.org/10.1038/s41477-018-0286-7

[72] Lescot, M., Piffanelli, P., Ciampi, A.Y., Ruiz, M., Blanc, G., Leebens-Mack, J., *et al.* (2008) Insights into the *Musa* Genome: Syntenic Relationships to Rice and Between *Musa* Species. *BMC Genomics*, **9**, Article No. 58. https://doi.org/10.1186/1471-2164-9-58

[73] Hoshino, A., Jayakumar, V., Nitasaka, E., Toyoda, A., Noguchi, H., Itoh, T., *et al.* (2016) Genome Sequence and Analysis of the Japanese Morning Glory *Ipomoea nil*. *Nature Communications*, **7**, Article No. 13295. https://doi.org/10.1038/ncomms13295

[74] Iorizzo, M., Senalik, D.A., Grzebelus, D., Bowman, M., Cavagnaro, P.F., Matvienko, *et al.* (2011) *De Novo* Assembly and Characterization of the Carrot Transcriptome Reveals Novel Genes, New Markers, and Genetic Diversity. *BMC Genomics*, **12**, Article No. 389. https://doi.org/10.1186/1471-2164-12-389

[75] Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., *et al.* (2016) A High-Quality Carrot Genome Assembly Provides New Insights into Carotenoid Accumulation and Asterid Genome Evolution. *Nature Genetics*, **48**, 657-666. https://doi.org/10.1038/ng.3565

[76] Mochida, K., Uehara-Yamaguchi, Y., Takahashi, F., Yoshida, T., Sakurai, T. and Shinozaki, K. (2013) Large-Scale Collection and Analysis of Full-Length cDNAs

from *Brachypodium distachyon* and Integration with *Pooideae* Sequence Resources. *PLoS ONE*, **8**, e75265. https://doi.org/10.1371/journal.pone.0075265

[77] Shalchian-Tabrizi, K., Minge, M.A., Espelund, M., Orr, R., Ruden, T., Jakobsen, K.S. and Cavalier-Smith, T. (2008) Multigene Phylogeny of *Choanozoa* and the Origin of Animals. *PLoS ONE*, **3**, e2098. https://doi.org/10.1371/journal.pone.0002098

[78] Tikhonenkov, D.V., Hehenberger, E., Esaulov, A.S., Belyakova, O.I., Mazei, Y.A., Mylnikov, A.P., *et al.* (2020) Insights into the Origin of Metazoan Multicellularity from Predatory Unicellular Relatives of Animals. *BMC Biology*, **18**, Article No. 39. https://doi.org/10.1186/s12915-020-0762-1

[79] Sakamoto, T., Deguchi, M., Brustolini, O.J., Santos, A.A., Silva, F.F. and Fontes, E.P. (2012) The Tomato RLK Superfamily: Phylogeny and Functional Predictions about the Role of the LRRII-RLK Subfamily in Antiviral Defense. *BMC Plant Biology*, **12**, Article No. 229. https://doi.org/10.1186/1471-2229-12-229

[80] Zorzatto, C., Machado, J.P., Lopes, K.V., Nascimento, K.J., Pereira, W.A., Brustolini, O.J., *et al.* (2015) NIK1-Mediated Translation Suppression Functions as a Plant Antiviral Immunity Mechanism. *Nature*, **520**, 679-682. https://doi.org/10.1038/nature14171

[81] Calil, I.P. and Fontes, E.P.B. (2017) Plant Immunity against Viruses: Antiviral Immune Receptors in Focus. *Annals of Botany*, **119**, 711-723. https://doi.org/10.1093/aob/mcw200

[82] Gouveia, B.C., Calil, I.P., Machado, J.P., Santos, A.A. and Fontes, E.P. (2017) Immune Receptors and Co-Receptors in Antiviral Innate Immunity in Plants. *Frontiers in Microbiology*, **7**, Article No. 2139. https://doi.org/10.3389/fmicb.2016.02139

[83] Teixeira, R.M., Ferreira, M.A., Raimundo, G.A.S., Loriato, V.A.P., Reis, P.A.B and Fontes, E.P.B. (2019) Virus Perception at the Cell Surface: Revisiting the Roles of Receptor-Like Kinases as Viral Pattern Recognition Receptors. *Molecular Plant Pathology*, **20**, 1196-1202. https://doi.org/10.1111/mpp.12816

[84] Zhang, X., Wang, X., Xu, K., Jiang, Z., Dong, K., Xie, X., *et al.* (2021) The Serine/Threonine/Tyrosine Kinase STY46 Defends against Hordeivirus Infection by Phosphorylating γb Protein. *Plant Physiology*, **186**, 715-730. https://doi.org/10.1093/plphys/kiab056

[85] Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) Database and Website. *British Journal of Cancer*, **91**, 355-358. https://doi.org/10.1038/sj.bjc.6601894

[86] Huang, P.J., Chiu, L.Y., Lee, C.C., Yeh, Y.M., Huang, K.Y., Chiu, C.H. and Tang, P. (2018) mSignatureDB: A Database for Deciphering Mutational Signatures in Human Cancers. *Nucleic Acids Research*, **46**, D964-D970. https://doi.org/10.1093/nar/gkx1133

[87] Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., *et al.* (2019) COSMIC: The Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research*, **47**, D941-D947. https://doi.org/10.1093/nar/gky1015

[88] Krassowski, M., Pellegrina, D., Mee, M.W., Fradet-Turcotte, A., Bhat, M. and Reimand, J. (2021) ActiveDriverDB: Interpreting Genetic Variation in Human and Cancer Genomes Using Post-Translational Modification Sites and Signaling Networks (2021 Update). *Frontiers in Cellular and Developmental Biology*, **9**, Article ID: 626821. https://doi.org/10.3389/fcell.2021.626821

[89] Wang, T., Ruan, S., Zhao, X., Shi, X., Teng, H., Zhong, J., *et al.* (2021) OncoVar: An Integrated Database and Analysis Platform for Oncogenic Driver Variants in Cancers. *Nucleic Acids Research*, **49**, D1289-D1301.

https://doi.org/10.1093/nar/gkaa1033

[90] Nakamura, Y., Komiyama, T., Furue, M., Gojobori, T. and Akiyama, Y. (2010) CIG-DB: The Database for Human or Mouse Immunoglobulin and T Cell Receptor Genes Available for Cancer Studies. *BMC Bioinformatics*, **11**, Article No. 398. https://doi.org/10.1186/1471-2105-11-398

[91] Gupta, S., Chaudhary, K., Dhanda, S.K., Kumar, R., Kumar, S., Sehgal, M., *et al.* (2016) A Platform for Designing Genome-Based Personalized Immunotherapy or Vaccine Against Cancer. *PLoS ONE*, **11**, e0166372. https://doi.org/10.1371/journal.pone.0166372

[92] Olsen, L.R., Tongchusak, S., Lin, H., Reinherz, E.L., Brusic, V. and Zhang, G.L. (2017) TANTIGEN: A Comprehensive Database of Tumor T Cell Antigens. *Cancer Immunology and Immunotherapy*, **66**, 731-735. https://doi.org/10.1007/s00262-017-1978-y

[93] Wu, J., Zhao, W., Zhou, B., Su, Z., Gu, X., Zhou, Z. and Chen, S. (2018) TSNAdb: A Database for Tumor-Specific Neoantigens from Immunogenomics Data Analysis. *Genomics Proteomics Bioinformatics*, **16**, 276-282. https://doi.org/10.1016/j.gpb.2018.06.003

[94] Xia, J., Bai, P., Fan, W., Li, Q., Li, Y., Wang, D., *et al.* (2021) NEPdb: A Database of T-cell Experimentally-Validated Neoantigens and Pan-Cancer Predicted Neoepitopes for Cancer Immunotherapy. *Frontiers in Immunology*, **12**, Article ID: 644637. https://doi.org/10.3389/fimmu.2021. 644637

[95] Zhang, G., Chitkushev, L., Olsen, L.R., Keskin, D.B. and Brusic, V. (2021) TANTIGEN 2.0: A Knowledge Base of Tumor T Cell Antigens and Epitopes. *BMC Bioinformatics*, **22**, Article No. 40. https://doi.org/10.1186/s12859-021-03962-7

[96] He, Y., Racz, R., Sayers, S., Lin, Y., Todd, T., Hur, J., *et al.* (2014) Updates on the Web-Based VIOLIN Vaccine Database and Analysis System. *Nucleic Acids Research*, **42**, D1124-D1132. https://doi.org/10.1093/nar/gkt1133

[97] Kim, P., Zhao, J., Lu, P. and Zhao, Z. (2017) mutLBSgeneDB: Mutated Ligand Binding Site Gene DataBase. *Nucleic Acids Research*, **45**, D256-D263. https://doi.org/10.1093/nar/gkw905

[98] Hu, R., Xu, H., Jia, P. and Zhao, Z. (2021) KinaseMD: Kinase Mutations and Drug Response Database. *Nucleic Acids Research*, **49**, D552-D561. https://doi.org/10.1093/nar/gkaa945

[99] Ryslik, G.A., Cheng, Y., Cheung, K.H., Bjornson, R.D., Zelterman, D., Modis, Y. and Zhao, H. (2014) A Spatial Simulation Approach to Account for Protein Structure when Identifying Non-Random Somatic Mutations. *BMC Bioinformatics*, **15**, Article No. 231. https://doi.org/10.1186/1471-2105-15-231

[100] Zhao, W., Yang, J., Wu, J., Cai, G., Zhang, Y., Haltom, J., *et al.* (2021) CanDriS: Posterior Profiling of Cancer-Driving Sites Based on Two-Component Evolutionary Model. *Briefings in Bioinformatics*, **22**, bbab131. https://doi.org/10.1093/bib/bbab131

[101] Garcia-Boronat, M., Diez-Rivero, C.M., Reinherz, E.L. and Reche, P.A. (2008) PVS: A Web Server for Protein Sequence Variability Analysis Tuned to Facilitate Conserved Epitope Discovery. *Nucleic Acids Research*, **36**, W35-W41. https://doi.org/10.1093/nar/gkn211

[102] Stranzl, T., Larsen, M.V., Lundegaard, C. and Nielsen, M. (2010) NetCTLpan: Pan-Specific MHC Class I Pathway Epitope Predictions. *Immunogenetics*, **62**, 357-368. https://doi.org/10.1007/s00251-010-0441-4

[103] Tung, C.W., Ziehm, M., Kämper, A., Kohlbacher, O. and Ho, S.Y. (2011) POPISK:

T-Cell Reactivity Prediction Using Support Vector Machines and String Kernels. *BMC Bioinformatics*, **12**, Article No. 446. https://doi.org/10.1186/1471-2105-12-446

[104] Oyarzún, P., Ellis, J.J., Bodén, M. and Kobe, B. (2013) PREDIVAC: CD4+ T-Cell Epitope Prediction for Vaccine Design that Covers 95% of HLA Class II DR Protein Diversity. *BMC Bioinformatics*, **14**, Article No. 52. https://doi.org/10.1186/1471-2105-14-52

[105] Dimitrov, I., Atanasova, M., Patronov, A., Flower, D.R. and Doytchinova, I. (2016) A Cohesive and Integrated Platform for Immunogenicity Prediction. In: Thomas, S., Ed., *Vaccine Design. Methods in Molecular Biology*, Humana, New York, 761-770. https://doi.org/10.1007/978-1-4939-3389-1_50

[106] Fleri, W., Paul, S., Dhanda, S.K., Mahajan, S., Xu, X., Peters, B. and Sette, A. (2017) The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. *Frontiers in Immunology*, **8**, Article No. 278. https://doi.org/10.3389/fimmu.2017.00278

[107] Zhang, S., Chen, J., Hong, P., Li, J., Tian, Y., Wu, Y. and Wang, S. (2020) PromPDD, a Web-Based Tool for the Prediction, Deciphering and Design of Promiscuous Peptides that Bind to HLA Class I Molecules. *Journal of Immunological Methods*, **476**, Article ID: 112685. https://doi.org/10.1016/j.jim.2019.112685

[108] Zhou, Z., Lyu, X., Wu, J., Yang, X., Wu, S., Zhou, J., *et al.* (2017) TSNAD: An Integrated Software for Cancer Somatic Mutation and Tumour-Specific Neoantigen Detection. *Royal Society Open Science*, **4**, Article ID: 170050. https://doi.org/10.1098/rsos.170050

[109] Popowicz, G.M., Müller, R., Noegel, A.A., Schleicher, M., Huber, R. and Holak, T.A. (2004) Molecular Structure of the Rod Domain of *Dictyostelium* Filamin. *Journal of Molecular Biology*, **342**, 1637-1646. https://doi.org/10.1016/j.jmb.2004.08.017

[110] Hsu, S.T., Cabrita, L.D., Fucini, P., Dobson, C.M. and Christodoulou, J. (2009) Structure, Dynamics and Folding of an Immunoglobulin Domain of the Gelation Factor (ABP-120) from *Dictyostelium discoideum*. *Journal of Molecular Biology*, **388**, 865-879. https://doi.org/10.1016/j.jmb.2009.02.063

[111] Hirose, S., Chen, G., Kuspa, A. and Shaulsky, G. (2017) The Polymorphic Proteins TgrB1 and TgrC1 Function as a Ligand-Receptor Pair in *Dictyostelium* Allorecognition. *Journal of Cell Science*, **130**, 4002-4012. https://doi.org/10.1242/jcs.208975

[112] Junqueira Alves, C., Yotoko, K., Zou, H. and Friedel, R.H. (2019) Origin and Evolution of Plexins, Semaphorins, and Met Receptor Tyrosine Kinases. *Science Reports*, **9**, Article No. 1970. https://doi.org/10.1038/s41598-019-38512-y

[113] Torruella, G., de Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M.A., del Campo, J., *et al.* (2015) Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Current Biology*, **25**, 2404-2410. https://doi.org/10.1016/j.cub.2015.07.053

[114] Burki, F., Kaplan, M., Tikhonenkov, D.V., Zlatogursky, V., Minh, B.Q., Radaykina, L.V., *et al.* (2016) Untangling the Early Diversification of Eukaryotes: A Phylogenomic Study of the Evolutionary Origins of *Centrohelida*, *Haptophyta* and *Cryptista*. *Proceedings. Biological Sciences*, **283**, Article ID: 20152802. https://doi.org/10.1098/rspb.2015.2802

[115] Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., *et al.* (2016) A New View of the Tree of Life, 2016. *Nature Microbiology*, **1**, Article No. 16048. https://doi.org/10.1038/nmicrobiol.2016.48

[116] Hehenberger, E., Tikhonenkov, D.V., Kolisko, M., Del Campo, J., Esaulov, A.S., Mylnikov, A.P. and Keeling, P.J. (2017) Novel Predators Reshape Holozoan Phylo-

geny and Reveal the Presence of a Two-Component Signaling System in the Ancestor of Animals. *Current Biology*, **27**, 2043-2050.
https://doi.org/10.1016/j.cub.2017.06.006

[117] Gitzendanner, M.A., Soltis, P.S., Wong, G.K., Ruhfel, B.R. and Soltis, D.E. (2018) Plastid Phylogenomic Analysis of Green Plants: A Billion Years of Evolutionary History. *American Journal of Botany*, **105**, 291-301.
https://doi.org/10.1002/ajb2.1048

[118] Cavalier-Smith, T. (2018) Kingdom *Chromista* and Its Eight Phyla: A New Synthesis Emphasising Periplastid Protein Targeting, Cytoskeletal and Periplastid Evolution, and Ancient Divergences. *Protoplasma*, **255**, 297-357.
https://doi.org/10.1007/s00709-017-1147-3

[119] Strassert, J.F.H., Jamy, M., Mylnikov, A.P., Tikhonenkov, D.V. and Burki, F. (2019) New Phylogenomic Analysis of the Enigmatic Phylum *Telonemia* further Resolves the Eukaryote Tree of Life. *Molecular Biology of Evolution*, **36**, 757-765.
https://doi.org/10.1093/molbev/msz012

[120] Li, L., Wang, S., Wang, H., Sahu, S.K., Marin, B., Li, H., *et al.* (2020) The Genome of *Prasinoderma coloniale* Unveils the Existence of a Third Phylum within Green Plants. *Nature Ecology and Evolution*, **4**, 1220-1231.
https://doi.org/10.1038/s41559-020-1221-7

[121] Federhen, S. (2012) The NCBI Taxonomy Database. *Nucleic Acids Research*, **40**, D136-D143. https://doi.org/10.1093/nar/gkr1178

[122] Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., *et al.* (2020) NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools. *Database*, **2020**, baaa062. https://doi.org/10.1093/database/baaa062

[123] Lang, B.F., O'Kelly, C., Nerad, T., Gray, M.W. and Burger, G. (2002) The Closest Unicellular Relatives of Animals. *Current Biology*, **12**, 1773-1778.
https://doi.org/10.1016/s0960-9822(02)01187-9

[124] Torruella, G., Derelle, R., Paps, J., Lang, B.F., Roger, A.J., Shalchian-Tabrizi, K. and Ruiz-Trillo, I. (2012) Phylogenetic Relationships within the *Opisthokonta* Based on Phylogenomic Analyses of Conserved Single-Copy Protein Domains. *Molecular Biology of Evolution*, **29**, 531-544. https://doi.org/10.1093/molbev/msr185

[125] Suga, H., Dacre, M., de Mendoza, A., Shalchian-Tabrizi, K., Manning, G. and Ruiz-Trillo, I. (2012) Genomic Survey of Premetazoans Shows Deep Conservation of Cytoplasmic Tyrosine Kinases and Multiple Radiations of Receptor Tyrosine Kinases. *Science Signaling*, **5**, ra35. https://doi.org/10.1126/scisignal.2002733

[126] Paps, J., Medina-Chacón, L.A., Marshall, W., Suga, H. and Ruiz-Trillo, I. (2013) Molecular Phylogeny of Unikonts: New Insights into the Position of Apusomonads and Ancyromonads and the Internal Relationships of Opisthokonts. *Protist*, **164**, 2-12. https://doi.org/10.1016/j.protis.2012.09.002

[127] Cao, L., Chen, F., Yang, X., Xu, W., Xie, J. and Yu, L. (2014) Phylogenetic Analysis of CDK and Cyclin Proteins in Premetazoan Lineages. *BMC Evolutionary Biology*, **14**, Article No. 10. https://doi.org/10.1186/1471-2148-14-10

[128] Sebé-Pedrós, A., Degnan, B.M. and Ruiz-Trillo, I. (2017) The Origin of *Metazoa*: A Unicellular Perspective. *Nature Reviews Genetics*, **18**, 498-512.
https://doi.org/10.1038/nrg.2017.21

[129] Naranjo-Ortiz, M.A. and Gabaldón, T. (2019) Fungal Evolution: Major Ecological Adaptations and Evolutionary Transitions. *Biological Reviews*, **94**, 1443-1476.
https://doi.org/10.1111/brv.12510

[130] Arroyo, A.S., Iannes, R., Bapteste, E. and Ruiz-Trillo, I. (2020) Gene Similarity Networks Unveil a Potential Novel Unicellular Group Closely Related to Animals from the Tara Oceans Expedition. *Genome Biology and Evolution*, **12**, 1664-1678. https://doi.org/10.1093/gbe/evaa117