Supplementary file 1 (accessible on www.papersatellitesjk.com)

to the paper: Antibody-like phosphorylation sites in focus of statistically based bilingual approach

Title:

More detailed descriptions of employed procedures and supplementary comments

WP1 Introductory information

WP1.1 Contents

Green titles denote here sections recommended to **first-rate attention of readers**.

WP2.3.4 Rules for co-localization and frequencies of MUSAS necessary for TCA

***

WP2.4 Final evaluation and re-evaluation of terminally displayed MPL segments

WP2.4.1 Steps of final MPL selection and determination of conserved domain context

WP2.4.2 Origin of seven alternative criteria determining terminal retrospective reselection

***

WP3 Comments to Figures

WP3.1 Time schedule of the main steps described in Figure 1

WP3.2 Combined statistical evaluation of graphs present in Figure 6

WP3.3 Smoothened histograms based on data-consistent Gaussian kernel measures

***

WP4 Comments to figure forms of tables

WP4.1 An approach to significance of sequence block columns determined in Figure 2

WP4.2 Subsets of the selected segments with distinct superior properties

WP4.2.1 Occurrence of hypermutation motifs

WP4.2.2 Domain and subdomain relationships

WP4.2.3 Extremes found during MRNS selection

WP4.3 MPL-like segments

***

WP5 Additional details and explanations

WP5.1 FFAS-derived sequence scanning (FFAS-scan)

WP5.2 A short overview concerning length equivalents

WP5.3 LE-derived fuzzy system

WP5.4 Protein queries determining nucleotide sequence queries MNSQ1 and MNSQ2 necessary for PPSIg-related database searches

WP5.5 Some details concerning MPQ3 restriction

WP5.6 Reasons for the usage of single-step BLASTN selection and combined cumulative approach based on two currently but differently adjusted TBLASTX

WP5.7 Additional notes on TCA

WP5.8 Additional trends in bioinformatic investigation of MPL

***

WP6. References to Supplementary File 1

WP1.2 Notes concerning the arrangement of Supplementary File 1

The chapters 1-4 of this file contain necessary supplementary information including all quantitative evaluations and main selection procedures important for all five steps of our approach (cf. Figure 1 and Figure 1B in SF3) and the following data processing. The last chapter (chapter five) comprises additional (not obligatory) information (i) enabling to better understand to procedures and published relationships or (ii) extending comments to the obtained results. Titles of all the following sections include WP

(abbreviation denoting web page) before order related numbers to distinguish these sections from those forming the paper.

WP1.3 Abbreviations in sections of web page supplement and in the text

For abbreviations see Supplementary File 2.

WP2 Methodological notes

WP2.1 Evaluation of sequence blocks and their segments

*WP2.1.1 BLAST-related length equivalents (**LE**).* LE-values determining F-values displayed in Figure 2 (see section 2.2.3 and Figure 3) were enumerated using PAM30 matrix (usually employed in BLASTP with chains of 5-15 aa) due to the limiting height of ten determined by the sequence block displayed in Figure 2. In accordance with the necessity to comprise evaluation of blank samples, we first of all expressed the formula for conditioned non-random Expect values ε(C,d,n,g) concerning each SBC:

$$\varepsilon(d,g,h,C) = E(C)/E(0) = \{K \times \varphi(d,g,h) \times \exp(-\lambda(C - \gamma(g))\}/\{K \times \varphi(d=1,g=0,h) \times \exp(-\lambda \times 0)\} =$$

$$\{\varphi(d,g,h)/\varphi(d=1,g=0,h)\} \times \exp\{-\lambda(C - \gamma(g))\} = \eta(d,g,h) \times \exp\{-\lambda(C - \gamma(g))\}, \quad \text{(WF1)}$$

where g is the number of gaps, h is the height of SBC (i.e. number of sequences composing evaluated MSA record); K and λ are constants (specifically K = 0.11; λ = 0.294); d is number of aa species (aa of different composition) within SBC; C is the sum of scores in SBC.

The quantities **φ(d,g,h), η(d,g,h), E(0), γ(g) and d** need more extended comments. The value **φ(d,g,h)** denotes the number of necessary operations required during SBC score enumeration approximated here by the formula:

$$\varphi(d,g,h) = sgn(g) \times h + d \times (h - g) = h \times \eta(d,g,h), \quad \text{(WF2)}$$

where g is the number of gaps; sgn(x) is function signum. In accordance with WF2, the value **η(d,g,h)** is equal to d in cases of non-gapped SBC and also in more general context η(d,g,h) refers to the extent of SBC diversity (cf. formula WF5). Random Expect (**E(0)**) represents in fact a necessary reference blank value when evaluating columns higher than one. In our approach E(0) = K x h, when considering random occurrence of aa (i.e. C = 0) and a model single step random choice in SBC without gaps, in accordance with WF1. This means that it holds φ(d=1,g= 0,h) = h, i.e. η(d,g,h) = φ(d,g,h)/h as in the last term of WF2.

Value **γ(g)** denotes SBC gap penalty, i.e. a value depending on the number of gaps **g**. In accordance with the fact that consistent maximum score or LE is always selected, g = n-1 gaps indicate more likely a single insertion than h - 1 deletions, i.e. such gap occurrence has to be penalized as a single “inverse” gap, when evaluating similarities **within SBC**. This means that more likely a number of penalized sites θ have to be specifically evaluated rather than a current the number of gaps g (cf. [1]). The following formula for gap penalties (**γ(g)**) assumes θ-value as consequence of approximately equal tendencies of diversifying protein sequences to insertion and deletions on aa level [2] (cf. also section 2.2.1):

$$\gamma(g) = 10 + \theta = 10 + Min(g,h-g) = 10 + h/2 - Abs(h/2 - g), \quad \text{(WF3)}$$

where Abs is an absolute value. In more general approach the number two in two denominators of last term of formula WF3 can be empirically specified, if possible. We have to add that parallel score

evaluation enumerates gaps as sites with score value of zero and specifically discriminates similarities of gaps as query structure (cf. WP2.4.1).

In accordance with our previous paper [3], the text above and section 2.2.1; score-related LE_i*, i.e. LE candidate alternative determined by i-th column aa, was enumerated by using Expect values ε described in formula WF1:

$$LE_i^* = \ln [\varepsilon_i(C_i,d,g,h)]/\ln[\varepsilon_i(0,1,1,s[x_i,x_i])] = \{\ln[\eta(d,g,h)] \times \exp [-\lambda \times (C_i - \gamma(g))]\}/\{ \ln[\eta(0,1,1)] \times \exp (-\lambda \times s[x_i,x_i])\} = \{\ln[\eta(d,g,h)] - \lambda \times [C_i - \gamma(g)]\}/\{1 \times (-\lambda \times s[x_i,x_i])\} = \{C_i - \gamma(g) - (1/\lambda) \times \ln [\eta(d,g,h)]\}/s[x_i,x_i]\} = \{S_i^{\circ} - \gamma(g)\}/s[x_i,x_i] = S_i/s[x_i,x_i], \quad \text{(WF4)}$$

where conditioned Expects $\varepsilon_i(C_i,d,g,h,)$ and $\varepsilon_i(s[x_i,x_i],0,1,1)$ correspond to SBC similarity with respect to i-th aa and compared i-th aa identity, respectively; S°_i and S_i are two types of SBC scores differently employed below and during enumeration of block similarity (cf. section 2.2.3); $s[x_i,x_i]$ denotes diagonal score of extended PAM30 matrix, i.e. score of i-th aa identity present in x_i-th row and column of this matrix (for extended PAM30 matrix see section WP2.1.4). In fact a single maximum LE_n (length equivalent) specifically represents each n-th SBC.

$$LE_n = \mathrm{Max}_{(i = 1 \text{ to } h)}\, LE_{in}^* = \mathrm{Max}_{(i = 1 \text{ to } h)}\, S_{in}/\mathrm{Abs}(s_n[x_i,x_i]) = \mathrm{Max}_{(i = 1 \text{ to } h)}\, \{C_{in} - [\gamma_n(g) + D_n]\}/\mathrm{Abs}(s_n[x_i,x_i]) = \{C_n - [\gamma_n(g) + D_n]\}/\mathrm{Abs}(s_n[x_i,x_i]) = S_n/\mathrm{Abs}(s_n[x_i,x_i]), \quad \text{(WF5)}$$

where $D_n = (1/\lambda) \times \ln [\eta_n(d,g,h)]$ is (in agreement with the preceding specification of η-value and the fact that in model case of g = 0 holds $\eta_n(d,g,h) = d$) the coefficient of diversity related to n-th SBC. This coefficient in fact reflects the current situation, when observing two SBC achieving the same C-value, but one of these SBC contains markedly higher number of different aa and thus appears to be more diversified. Due to the negative value in diagonal matrix, an element evaluating X-X "identities", $\mathrm{Abs}\{s_n[x_i,x_i]\}$ substitutes $s_n[x_i,x_i]$ in formula WF5 (like in formula 1 of the paper text) to indeed select actual maximum.

In the non-invasive approach described in this paper, we evaluated only aa present in the corresponding SBC. To enumerate LE_i-related to absent aa in case of "invasive" approach, we propose to substitute the original **d-value** by d* = d+1 (cf. formula WF2), if aa different from those present in SBC achieves maximum LE. Such comparison may have some reason only in case of SBC with low or negative LE values.

Similarly to formulas WF4 and WF5 we can also define σ-consensus-related values S_n* and S_n°* (cf. section 2.2.1):

$$S_n^* = \mathrm{Max}_{(i = 1 \text{ to } h)}\, S_{in} = \{\mathrm{Max}_{(i = 1 \text{ to } h)}\, S_{in}^{\circ}\} - \gamma_n(g) = C_n^* - D_n - \gamma_n(g) = S_n^{\circ *} - \gamma_n(g), \quad \text{(WF6)}$$

where the value S_n°* represents maximum column-related score which necessary for evaluation of SBC of overall sequence block (or block segment) similarities described in section 2.2.3.

Due to different $s_n[x_i,x_i]$ values, different aa in n-th SBC (i.e. different i-values in formulas WF5 and WF6) can sometimes determine different values of LE-related S_n and score maxima S_n*. In accordance with this fact and under conditions of low S_{in} values (lower than current statistically derived limiting values), maximum SBC score S_n* can be accompanied by aa with high score identity $s_n[x_m,x_m]$ determining together with S_n* value absence of required LE-derived CBS similarity (cf. Figure 3). On the other hand, S_{in} values can unexpectedly achieve required level for LE-derived similarity in the same SBC

due to lower $s_n[x_i,x_i]$ value, and we thus indicate contradiction with respect to presence of similarity. This rare disagreement led us to propose procedure comprising both S_n and S_n* during two-step evaluation. More precisely, we first of all restrict CSB as complex of individual SBC based on S_n-related LE-derived F-values (cf. sections 2.2.2, 2.2.3, WP5.3 and Figure 3). In the second step, we then verified given restriction of CSB using conventional approach based on maximum SBC score S_n* and the corresponding optimized evaluation of gaps (cf. section WP2.1.2). The preceding considerations in fact demonstrate limited but existing **consensus duality** mentioned in section 2.2 and indicated by Greek alphabet in Figure 2. Further research will perhaps brings mathematical unification of these procedures, if the corresponding solution would not considerably increase time of enumeration.

WP2.1.2 Optimal gap-related matrices necessary for sequence block evaluation. Let G[j,k] be a map of gap positions, where 1 and 0 indicate the presence and absence of gaps, respectively. Let P[j,k] be a map of penalized sites optimized with respect to random SBC-related insertions and deletions, i.e. for each individual m holds: if $g[m] \leq (h/2)$ then P[j,m] = G[j,m], else P[j,m] = 1 - G[j,m] (where $g[m] = \Sigma_{j=1 \text{ to } h} G[j,m]$). Let Q[j,k] be a map of insertion-related sites inverse to gaps, i.e. for each individual m we can write Q[j,m] = sign (g[m]) x (1 - G[j,m]). In accordance with the preceding definitions of matrices, we can enumerate the value of optimized gap penalty **Θ(X*[j,k])** of any oblong or square sequence block:

$$\Theta(X^*[j,k]) = \text{Min}\{\Theta(G[j,k]), \Theta(P[j,k]), \Theta(Q[j,k])\} = \text{Min}_{t = 1 \text{ to } 3}\{\Theta\{X_t[j,k]\}, \quad \text{(WF7)}$$

where $X_n[i,k]$ are the evaluated matrices, Min is minimum of the introduced $\Theta\{X_t[j,k]\}$ values; X*[j,k] is selected optimized matrix. The $\Theta\{X_t[j,k]\}$ values were enumerated using the formula:

$$\Theta(X_t[j,k]) = 10 \text{ x } \{\Sigma_{j=1 \text{ to } h} \text{sign}(\Sigma_{k=1 \text{ to } N} X_t[j,k])\} + \Sigma_{j=1 \text{ to } h} \Sigma_{k=1 \text{ to } N} X_t[j,k], \quad \text{(WF8)}$$

where N is number of columns composing evaluated sequence block.

Formula WF8 determined an optimized number r of X*[j,k] rows including positive values and adequate topical γ_k values:

$$r = \Sigma_{j=1 \text{ to } h} \text{sign}(\Sigma_{k=1 \text{ to } N} X^*[j,k]), \quad \text{(WF9)}$$

$$\gamma_k = \Sigma_{j=1 \text{ to } h} X^*[j,k]. \quad \text{(WF10)}$$

In case of CSB1 and CSB2 evaluated in our paper and frequently also in other cases of CSB, it holds: X*[j,k] = G[j,k] = P[j,k]. This means that the number of chains with gaps and the numbers of gaps in individual SBC are immediately enumerated as n and γ_k, respectively, when determining overall penalty for gaps in a sequence block (the simplest situation equal to current BLAST evaluation).

The separate evaluations of SBC based on LE or σ_k follow from P[j,k] and represent in fact local independent one-dimensional evaluation. The differences between overall block and local SBC evaluations thus resemble differences between total and partial derivations.

WP2.1.3 LE-derived restriction of highly conserved islands. In case of absence of sequence pattern necessary for restriction of conserved block segment, we propose to evaluate presence of quasi-conserved segments named here as highly conserved islands (**hci**). Our first attempt to define hci includes restrictions of short sequence block segments with SBC simultaneously limited by two parameters, i.e. j and k:

$$j = \text{int}\{(n-1)/3\} - 1, \quad \text{(WF11)}$$

$$k = \text{int}(n/5) - 1, \quad \text{(WF12)}$$

where j is the maximum number of SBC with F values: $0 \leq F < 6$; k denotes maximum number of SBC with $F < 0$; int denotes an integer value composing the given number; n is the overall number of SBC in hci. Restricted hci can be among others used in PHI BLAST searches as alternative though usually shorter, sources of queries than the conserved regions described in 2.2.3.

WP2.1.4 Simple usage of MSA-record-related pointer matrix in sequence block evaluation. Each single character determining aa or a group of aa (currently used in substitution matrices) represented specifically one of 23 numbers indicating both column and row numbers of a symmetrical substitution matrix (in our case PAM30). In addition, character J=24 substituted gaps and pointed thus to rows and columns with zero in an extended substitution matrix (e.g. **MatS** employed below). This matrix was loaded during the run of **initial subprogram** together with the number code necessary for the following translation of protein sequence characters to the Lists containing numbers pointing to the aa-related position orders (corresponding to rows and columns) in MatS. In accordance with this code, dashes present in initial MSA-record were rewritten to sequences with J substituting gaps. Subsequently, commas were inserted among characters, closing the modified sequence in parentheses. All these steps were necessary for input procedures (e.g.: {S,V,T,L,J,C,S} → List4) of the **second subprogram**, translating sequences to numbers pointing to positions in MatS. Keeping the orders of the corresponding chains in evaluated MSA-record, when loading the corresponding sequence-related Lists to a matrix during the **third subprogram**, we assembled matrix **MatA** containing numbers unambiguously describing original MSA record. In case of BASIC language **transposed form of MSA-record-related matrix** MatA (represented below by the resulting matrix **MatAns**) was necessary for evaluation of MSA-record-related columns. During this evaluation (**fourth subprogram**), each number present in MatAns was separately and continuously combined with all numbers present in the same row of the same matrix. This combination of numbers then determined the rows and columns in substitution matrix MatS containing the relevant scores of the corresponding separate aa comparisons. In addition, these numbers also determined frequencies and positions of gaps in separate SBC.

In accordance with the displayed code, List15 enables the restriction of conserved block segments (**CSB**; see sections 2.2.2 and 2.2.3), whereas Lists 16 and 19 together with MatG are necessary for CSB enumeration (see also sections 2.2.3 and WP2.1.2). Processing of MatB and MatC then yields the composition of σ- and LE-related consensi, respectively (cf. section 2.2 and sections of WP2.1).

WP2.2 Construction of two multi-nucleotide-sequence queries (MNSQ) employed in BLAST searches for PPSIg-related segments

WP2.2.1 Overall scheme of MNSQ formation. An overall scheme including processes forming both MNSQ is displayed in Figure 1C present in the supplementary file SF3.

*WP2.2.2 Construction of hybrid query sequences (**HQS**) based on selective substitution of typical sequences in sites co-localizing with well-correlated consensus amino acids.* HQS were generated as queries necessary for the second variant of initial searches described in the supplementary file SF3. Since aa of **consensus 1** (CSB1-related consensus equal to both σ- and LE-consensi) achieving SBC of $F \geq 6$ attained simultaneously previously described limit $LE \geq 3$, all these double-positive aa were able to

substitute co-localizing CSB1-chain-related aa in agreement with the procedures mentioned in the previous paper [4]. Parallel profile-related statistical restriction of possible substituting aa in consensus 1 (i.e. $p < 0.005$ [1,5]; see also 2.1.2) enabled also a similar generation of **HQS**. However, these substitutions were low frequent, i.e. only poorly efficient with respect to the required minimum sequence difference between differently generated HQS pairs or pairs composed of a typical sequence and HQS (difference in at least two aa was required). Consequently, only LE-derived HQS more different from chains of CSB1 were accepted as HQS useful for selection of List2o1 (see Figure 2, section WP5.4 and supplementary file SF3).

WP2.2.3 Steps restricting initial item lists of clone names (L[s]o1). TBLASTN searches with individual protein QS, i.e. Q{q[s]}, determined initial search records (Records 1):

TBLASTN{s; Q{q[s]}; WMR:W3B62; if s=1, then SR:*Elasmobranchii*[ORGN] else (s=2)

SR:None[ORGN];E≤10^-4;S≥40bits}→Records1[s,q[s],i], (WF13)

where s=1 and s=2 indicate strategies necessary for initial restriction of MNSQ1 and MNSQ2, respectively; q[s] are QS numbers; i is the order of the selected sequence item; WMR is search restriction concerning word size and matrix (W3B62 denotes current BLAST adjustment including word size 3 and matrix BLOSUM62); SR means species restriction of the search; E denotes Expect value; S is bit score. For the familiar group relationship of Expect and score limits see section 2.2.3.

The following processing of Records1[s,q[s],i] comprised several selections, anti-redundant procedures, species-related randomization and iterative filling/refilling of items. This processing can be illustrated by the following formulas:

Records1[s,q[s],i] → Records2[q[s],j] → Records3a[s,q[s],k,t=n] → Records3b[s,f,g,h,m≤M,t=n] → Records3c[s,f,g,h,m≤M-y[g,h,t],t=n]→ If $(\sum_{g=1 \text{ to } M} \sum_{h=1 \text{ to } M} y[g,h,t]>0)$, then Records3c[s,f,g,h,m≤M-y[g,h,t],t=n] → incompletely filled Records3a[q[s],k≤M-y{[g,h,t]→q[s]},t=n+1→n], else Records3c[s,f,g,h,m≤M,t=n] → Records3d[s,f,g,h,m], (WF14)

Records3d[s,f,g,h,m] → L[s]o1, (WF15)

j,k,m are transposed numbers determining sequence item orders in the corresponding sets; **f** represents the main feature indicator associated with rearrangement of Records3 {set with f=1 (i.e. set **SIMULT**, unique for each s-value and unlimited with respect to species origin and sample capacity) including sequence items redirected from the sets originally achieving f=2 (see below); the set of items with f=2 was selected using **STS**-queries (segments of typical sequences forming CSB1 and used as queries in the intial searches) or their hybrid derivatives with consensus (cf. WP2.2.2); the items of sets with f=3 and f=4 were found using experimentally derived consensus-like sequences cls (only if s=2) or with the CSB1-related consensus 1, respectively (cf. sections 2.1.1, 2.3.1 and Figure 2)}; **g** denotes the number of QS groups, each represented by (i) STS-queries only (if s=1 then it always holds h=1) or (ii) pairs of sequences in case of s=2, where STS-query and its hybrid derivative (HQS) determine hierarchy levels h=1 and h=2, respectively; **t** is the number of iteration steps comprising (i) primary filling or refilling and (ii) diminishing the redundancy (see below); **M** is usually equal to maximum Records3 capacity (see below), i.e. M=M[s,t], except for rare ineffective searches determining M = M[g,h,t] = J < M[s,t], where J denotes the order of last j-th items of Records2 achieving sufficient score and Expect value; **y[g,h,t]** and

y{q[s],t} are the numbers of empty sites after t-th revision of sequence item redundancy. The existence of only two initial lists (L1o1 and L1o2) followed from the identity of σ- and LE-related consensi, i.e. unique consensus 1, derived from the topical CSB1.

The **first transition** (**first dart**) in the formula WF14 consisted in: (i) removing items with sequence shorter than 100 nucleotides and longer than 3000 nucleotides, (ii) selection of TBLASTN similarities containing the pattern CX(10,13)WXXQXP (majority of items), and (iii) revision of sequence redundancy in individual search records. The **second transition** represented species limited filling or refilling of items to Records3a-related sets with f>1 achieving only limited capacity of the sites for the items (see below). This filling/refilling occurred mostly gradually in accordance with increasing size of Expect values, except for the specific preference of complete or standardized sequences in cases of indicated redundancies. Maximum number of inputted non-redundant items of the same species origin was always five for each set of Records3. The difference in maximum **capacity** of two Records3[s,f >1] followed from both (i) a number of QS and (ii) extent of database determining the generation of MNSQ1 and MNSQ2 units, respectively. Consequently, at most ten and twenty items were accepted to each set of Records3[s=1;f>1] and Records3[s=2;f>1], respectively. **The third transition** yielded a useful rearrangement of Records3 enabling subsequent oriented elimination. During this step (denoted by fourth dart) all items related to certain s-value but repeating in groups characterized by f=2 and different g-values were first of all united and redirected to an unlimited set of preferred sequence items SIMULT (f=1). **Subsequently**, all other redundant sequence items were removed, keeping only sole records with minimum f or minimum h, if f value was the same. Newly appearing empty sites were shifted to the end of each set. The resulting Recods3c were then rearranged to Records3a and refilled. Alternative absence of empty sites in Records3c indicated the terminal stage of record Records3 (Records3d[s]) determining two output lists (formula WF15; for form of individual lists see section WP2.3.1).

WP2.2.4 Revision of the selected lists (supersets) of items and formation of both MNSQ. The re-selection included three pairs of subsequent searches with different multi-sequence protein queries containing consensus 1 (**MPQ1**) and LE consensus segments MPQ2a, MPQ3a grading as MPQ[u] with the size of u-value (u = 1 to 3), regarding s and x (s = 1 or 2 and x = 1 or 2).

TBLASTN{s; MPQ[u]; if WMR:W2P30, then x=1; else (WMR:W3B62) x=2; ER via List: L[s]o[u];
E ≤ 10^-4; S ≥ 40 bits} → r-Records[s,u,x,z]→ List[s,u+1,x], (WF16)

Lists[s,u+1,x=1] ∩ Lists[s,u+1,x=2] → L[s]o[u+1→u], (WF17)

where W2P30, W3B62 are two currently but differently adjusted BLAST using word size 2 plus matrix PAM30 or word size 3 and matrix BLOSUM62, respectively; ER is Entrez restriction; initiating L[s]o1 was derived in formula WF15 (section WP2.2.3); r-Records are re-selection records; x enumerates WMR alternatives; z encodes item orders. For the reason of convention forming List[s]o[u] see supplementary file SF3.

The following second revision (**R2**) with MPQ2b and MPQ3b (containing σ-consensus segments instead of LE-derived ones) included also variation with respect to x (x =1 or 2; cf. formulas WF16 and WF17). R2 was initiated by comparison of MPQ2b with the list L[s]o4. Instead of positive selection used in the first revision step, R2 identified items failing to achieve Expect and score limits common for all

selection procedures generating MNSQ. R2 resulted in a list of a single item (L[s]o5neg). This list was subtracted from conjunction of r-Records[s,3,x=2,z] and L[s]o4, determining thus final records f-Records[s,z]. In accordance with the following formula all operations were finished with an assembly of MNSQ:

(r-Records[s,u=3,x=2,z] ∩ (List[s]o4) NOT L[s]o5_neg → f-Records[s,z] → MNSQ[s], (WF18)

where the last dart denotes (i) online translation of TBLASTN records to nucleotide sequences using ALIGN (BLAST-2 sequences), (ii) removal of residual sequence redundancy, (iii) formation of MNSQ[s] units and (iv) fusion of these units into a single sequence [6]. In fact each selected oligonucleotide sequence segment formed a sequence unit of 150-nucleotides length together with the separating sequence spacers composed only of repeating characters N only.

WP2.3 BLAST searches using MNSQ1 or MNSQ2 and accompanying procedures for selecting PPSIg-related nucleotide sequences

*WP2.3.1 A more detailed description of Entrez restrictions (**ER**) necessary for all searches with MNSQ1 or MNSQ2.* In accordance with Figure 1, five ER strategies determined: (i) molecules selected by FFAS scan (sections 2.1 and WP5.1) including close fold-related paralogues of antigen receptors (formulas WF19 and WF20), (ii) molecules of superior MNSQ1 and MNSQ2 similarities outside familiar groups of antigen receptors (formula WF21); (iii) oncogenes and molecules involved in phosphorylation and dephosphorylation (formula WF22) and (iv) molecules including multiple cancer names in their title (formula WF23; the list was performed in accordance with the book dealing with oncogenes and cancerogenesis [7]:

AY238517.1 OR NM_008798.2 OR NM_147130.2 OR AY457047.1 OR NM_007261.3 OR NM_001289085.1 OR AK298752.1 OR NM_139018.4 OR NM_001289084.1 OR NR_110298.1 OR AY039664.1 OR NM_001768.6 OR NM_001081110.2 OR BE348845.1 OR BC053400.1 OR NM_134248.2 OR AF508193.1 OR AY078502.1 OR BC043216.2 OR BC048589.1 OR DQ087183.1 OR NM_182607 OR NM_026103.1, **(WF19)**

(Homo sapiens[ORGN] OR Mus musculus[ORGN]) AND (programmed cell death protein 1[TI] OR NCR3[TI] OR natural cytotoxicity triggering receptor 3[TI] OR CD300[TI] OR CD300*[TI] OR CLM-1[TI] OR CD300a[TI] OR CD300lf[TI] OR CD300LF[TI] OR CMRF35-H[TI] OR CMRF35-like*[TI] OR CLM1[TI] OR CMRF35H[TI] OR CMRF-35-like*[TI] OR CD8 alpha[TI] OR CD8a[TI] OR (CD8[TI] AND alpha[TI]) OR (hepatitis A[TI] AND virus[TI] AND receptor[TI]) OR trem-like[TI] OR TREML1[TI] OR TLT1[TI] OR ((V-set Ig domain[TI] OR V-set immunoglobulin domain[TI]) AND containing protein[TI])) AND (cDNA[TI] OR mRNA[TI]), **(WF20)**

(mRNA[TI] OR cDNA[TI]) NOT (immunoglobulin[TI] OR T-cell receptor[TI] OR T cell receptor[TI] OR TCR*[TI] OR TCRV*[TI] OR IgA[TI] OR IgD[TI] OR IgE[TI] OR IgG[TI] OR IgM[TI] OR kappa[TI] OR lambda[TI] OR variable[TI] OR heavy chain[TI] OR IgV*[TI] OR IgH*[TI] OR light

chain[TW] OR IgL[TI] OR antibody[TI] OR nanobody[TI] OR monoclonal[TI] OR hybridoma[TI] OR VH*[TI]), **(WF21)**

(Homo sapiens[ORGN] OR Mus musculus[ORGN]) AND (kinase[TI] OR phosphatase[TI] OR phosphorylase[TI] OR oncogene[TI]) AND (mRNA[TI] OR cDNA[TI]), **(WF22)**

(Mus musculus[ORGN] OR Homo sapiens[ORGN]) AND (cancer[TI] OR leukemia[TI] OR lymphoma[TI] OR sarcoma[TI] OR carcinoma[TI] OR neuroblastoma[TI] OR glioblastoma[TI] OR retinoblastoma[TI] OR melanoma[TI] OR pheochromocytoma[TI] OR osteosarcoma[TI] OR fibrosarcoma[TI] OR chondrosarcoma[TI] OR rhabdomyosarcoma[TI] OR hemangiosarcoma[TI] OR liposarcoma[TI] OR myeloma[TI] OR myosarcoma[TI] OR lymphoblastoma[TI]) AND ((mRNA[TI] OR cDNA[TI]) NOT (immunoglobulin[TI] OR Ig[TI] OR antibody[TI] OR IgM[TI] OR IgG[TI] OR IgA[TI] OR IgH[TI])). **(WF23)**

The use of ER strategies including Boolean operator "NOT" appeared to be more complicated in the last edition of BLAST in contrast to older versions. Consequently, two following additional rules concerned formula WF21. (i) Species specification of WF21 was performed using species restriction but not via ER. (ii) Since the search records obtained with the help of WF21 were unique ones, which did not contain additional functional or phylogenic context, we did not further process sequence items with uncertain title name, such as unnamed, unknown, similar, etc. In addition, some sequence items with less uncertain topological names or sequences achieving a score of at least 50 bits in any record of searches with MNSQ were revised with respect to their identity with antigen receptors (**AR**) using (i) revising BLASTN searches with whole sequences or (ii) notes in databases confirming antigen receptor nature of the corresponding sequences. In accordance with the former alternative and current bit-score-derived classification of BLAST similarities, the molecules achieving similarities at least 200 bits in their revising BLASTN searches were identified as molecules of discriminating familiar relationship to AR.

As follows from formulas WF19 and WF20, searches for paralogues were limited by their occurrence in human and mouse genomes except for searches for novel immune-type receptors (NITR10 and NITR11) selected by FFASscan because NITRs are not present in mammalian genomes. Since a familiar group of NITR genes belong to the phylogenically closest relatives to AR [8,9], NITR represent important reference molecules. Consequently, the searches for NITR10 a 11 were unique in that they did not undergo species restriction. This implicated the following strategy for NITR:

(novel immune-type receptor 10[TI] OR novel immune-type receptor 11[TI] OR NITR10[TI] OR NITR11[TI]) AND (cDNA[TI] OR mRNA[TI]). **(WF24)**

In addition, similarly to global searches for superior similarities (cf. WF21), generalized searches for superior similarities concerned all NITR. These searches were limited by a minimum score of 35 bits and included BLASTN derived items and double restricted TBLASTX items (cf. section 2.4). The strategy was similar to the preceding one:

(novel immune-type receptor[TI] OR NITR*[TI]) AND (mRNA[TI] OR cDNA[TI]). **(WF25)**

WP2.3.2 Regularity of the bit score limits for MRNS-related BLASTN, TBLASTX and ternary combined (TCA) approaches. We currently distinguish two types of short sequence similarities achieving lengths related to subsequently observed PPS, i.e. pattern-related similarities (**PS**; lengths **3-10 aa/7-30 nucleotides and rarely also 2 aa**) and similarities classified as short nearly exact matches (**SNEM**; lengths **5-15 aa; 15-45 nucleotides**).

Three important and frequent cases can be seen, when evaluating PS, i.e. (i) occurrence of unambiguously determined sequence **patterns in a representative MSA-record** (e.g. three or more near SBC evaluated by $p < 10^{-6}$ and containing solely common aa represent very favorable arrangements; cf. WP4.1), (ii) unexpectedly **high** pattern **frequencies** (significantly increased numbers with respect to expected or actually compared valid reference values; $p < 0.05$) in wide-ranging or specific sequence databases and objectively restricted compared sequence database subsets, (iii) experimentally verified functionally important patterns (case of functional motifs). In all three given pattern cases, evaluation of PS validity can be performed independently of database length. Hence non-context (extracted) pattern sequence similarities are evaluated only based on numbers and species of pattern aa and if need be according to length variability of patterns. Such evaluation can be among others approximated by minimum limiting bit scores (S*) as pattern limits in case of BLAST evaluation under the simplified conditions of not fully identical sequences. The corresponding interval pattern limits are then enumerated using the formulas:

$$S1^* = (1/\ln 2) \times \ln \{[Min(M) \times Min(N)]/Max(E)\}, \quad \text{(WF26)}$$

$$S2^* = (1/\ln 2) \times \ln \{[Max(M) \times Max(N))/Max(E)\}, \quad \text{(WF27)}$$

where S1* and S2* are bottom and upper limiting score values; M and N are lengths of the whole query (MNSQ1 or MNSQ2) and the compared subject segments (i.e. MRNS), respectively. The described usage of limiting Expect values (Max(E)) instead of w (determining $p < w$) follows from the fact that $p < E$ and $p \approx E$, in cases of $E \leq 0.05$. The value Max(E) is selected according to the type of evaluation (it is frequently 0.05, 0.01 or 0.005; for our specific solution see below).

The formulas WF26 and WF27 were derived from BLAST formulas described below. More precisely, the formula WF28 and the first term of WF30 represent widely known BLAST formulas, whereas WF29 followed from WF28. We distinguished here score (S) and bit score (S*):

$$E = K \times M \times N \times \exp(-\lambda \times S), \quad \text{(WF28)}$$

$$S = (1/\lambda) \times \ln \{(K \times M \times N)/E\}, \quad \text{(WF29)}$$

$$S^* = (1/\ln 2) \times \{\lambda \times S - \ln (K)\} = (1/\ln 2) \times \{\ln [(K \times M \times N)/E] - \ln(K)\} = (1/\ln 2) \times \ln\{(M \times N)/E\}, \quad \text{(WF30)}$$

Current BLAST limits are mostly too strict to found structurally important short dense sequence similarities. Nevertheless, the specific **SNEM**-related parameters are automatically adjusted during BLAST comparison, when inputting short sequence segments as queries of our interest. The ratios between current BLAST and SNEM limits are R = 100 and R* = 2000 in cases of BLASTN and BLASTX, respectively. In accordance with the employed bilingual approach (evaluating both nucleotide and protein sequence similarities), due to lower R-value (i.e. the value yielding stricter limit) BLASTN

appears to be crucial for transformation of standard BLAST-related bit score limit to its regular SNEM-related counterpart.

The standard bit score limit of BLAST, i.e. Min(S*) = 38 bits, follows from reference marginally significant cases $p < 0.05$, related to extended BLAST similarities [1]. This score limit is closely related to (and perhaps represents non-rounded version of) the bottom limit for middle-range BLAST similarities of 40 bits present in all headings of BLAST records. To sufficiently transform Min(S*) to the corresponding SNEM value, it is possible to employ several (at least four) local logarithmic-linear dependences of bit score on logarithms of Expect values differently based on several corresponding BLASTN records and the following formula:

$$S^*[E] = a \times \log(E) + b, \qquad \text{(WF31)}$$

where a and b are constant values in the selected discrete interval; a-value is negative, whereas b-value is positive.

Based on WF31-related log-linear graphs, we can assess various (**search-dependent**) values of Min(S*)-related Expects for the critical score, i.e. several E[38]. Subsequently, it is possible to calculate transformed SNEM-related Expect limits Max**E(R,S*) using R = 100 mentioned above:

$$Max^{**}E(R,38) = R \times E[38], \qquad \text{(WF32)}$$

In the following step, final almost the same (i.e. **search-independent**) SNEM score limits **S**[R,38]** corresponding to different values E = Max**E(R,38) can be determined when using inversely and specifically local log-linear graphs of S*[E] described in formula WF31 (for actual step compensating errors in employed approximations see below).

The sequence comparisons yielding MRNS evaluated in our paper (cf. section 2.4) determine in fact the **third type of similarity**, looking like a **hybrid of the two types of similarities mentioned above**. Hence, pre-formed double selected MNSQ units (determining MRNS) were derived as actual top homologues to (i) sequences of standardized conserved sequences (representing consensi generated before alignment and specified by CDD-search of BLAST based on sophisticated rules) or (ii) other typical sequences selected via sharply limited procedures (igw and cls; cf. sections 2.1.1, WP5.4, Figure 2 and the corresponding paper concerning to IgV evalution [4]). Most of MRNS lengths moreover occur in the range of 5-10 aa/15-30 nucleotides sufficient for both PS- and SNEM-related length restrictions. Consequently, the described hybrid properties **together with** the more specific fact that we compare query sequences of **high variability** (IgV-related sequences) led us to approximating the bottom bit score limit for MRNS searches (**B**):

$$B = 0.5 \times (S^{**} + \mathrm{Max}(S1^*, S2^*)). \qquad \text{(WF33)}$$

In accordance with formula WF33, B allows the usage of scores occurring only in a **guaranteed upper half of the twilight zone** between the limits for PS and SNEM.

In the **three-step enumeration**, we selected or determined all necessary values. The Max(E) = **0.005** value, employed in the first step, was in agreement with the limit $p < 0.005$ (indeed 5×10^{-3}) restricting PSI BLAST evaluating multi-sequence relationships (cf. section 2.2.3 and [1]). The usage of this PSI BLAST limit was coherent with multi-sequence repetition of MNSQ units substituting in fact a unique motif in our case. Based on Max(E), other necessary input values (Min(M) = 6150 (length of MNSQ1),

Min(N) = 15, Max(M) = 16350 (length of MNSQ2), Max(N) = 39) and formulas WF26 and WF27, we restricted the interval of bit score limits for PS, i.e. S1* = **24.14** and S2* = **26.93**. In the second step, several local log-linear dependences obtained from different BLAST records enabled us to approximate SNEM bit score limit S**[R,38] = **31.9** with a low error of the order of 10^-2. Based on S1*, S2* and S**, we enumerated the bottom bit score limit for MNRS searches B = **29.4**. Finally, we estimated **the limit of thirty bits as regular bit score minimum** in our searches for MRNS similarities in the **two positively Entrez-restricted BLAST approaches** (single step BLASTN and double restricted TBLASTX searches) reflecting the attention focused on **cancer-** and **phosphorylation-related molecules** (cf. Introduction and section 2.4). This final substitution of B was performed, when considering (i) rounding off conventions currently used in similar estimations, (ii) actual row of discrete bit scores values near B-value observed in the BLASTN records enabling the score limit transformation (bit scores 28.3, 30.1, 31.9) and (iii) possible errors due to the employed approximations.

In accordance with the rounding off conventions mentioned above, the maximum score limit of **25 bits for TCA searches** was postulted in accordance with the values restricting the position of enumerated interval of PS limits. The limit 22 bits for MUSAS was then estimated in accordance with the limit for distant phylogenic relationships derived in one of our papers [4]. The possibility of such limit minimization followed from the presence of two additional structurally important demands, i.e. (i) repeated occurrence of the same subject sequences in the limited similarities derived by two different searches and (ii) achievement of frequency limit for MUSAS in these two searches (MUSAS are repeating similarities with different MNSQ units with almost the same subject sequence positions; cf. WP2.3.4). In consequence of the described ternary restriction, the numbers of MRNS obtained with TCA were still about two times lower than in the case of one-step (and less diversified) BLASTN searches limited by 30 bits. Consequently, we assume that TCA appears to be (in accordance with their frequency as **a posteriori experience**) an even more selective and thus more regular approach (with respect to the number of finally displayed MPL) than the compared BLASTN searches, which were restricted above as regular searches.

Similarly to the five-bit-score difference between group-related maximum-bit-score limits for (i) the positively Entrez-restricted BLASTN searches and (ii) TCA of the same restriction, the third alternative value of limiting maximum-bit-score employed in our paper differed also by five bits. This score value (35 bits) limited score maxima in the general approach. This approach comprised non-focused negatively Entrez-restricted BLASTN searches, i.e. searches for any sequences different from antigen receptors. In fact, increased superior bit score limit minimized the obtained in part complementary sequence sets to executable extent corresponding to the introductory stage of the described methodological approach. In summary, we thus postulated here useful and adequate **five-bit-related grading** of regular maximum bit score limits comprising the values of **25**, **30** and **35** bits. For details concerning Entrez-restriction strategies dealt with here see section WP2.3.1.

WP2.3.3 Paths of the BLAST searches composing TCA. Since all thousandth items of initial search records of TCA achieved a score substantially higher than (i.e. far from) the limiting bottom value of "maximum scores" of 25 bits, the composition of final records depended on the order of the search types.

This means that differently arranged paths including the same searches achieved at least partially different results. TCA differing in their ER (cf. WF22 and 23) and compared MNSQ underwent the same search paths:

N → X2 →N, (WF34)

N → X3 → N, (WF35)

X2 → X3 → X2, (WF36)

X3 → X2 → X3, (WF37)

where the first, second and third components of the formulas represent records of initial complementary and reverse searches described in WP5.7, respectively; N denotes BLASTN records; X2 and X3 represent TBLASTX searches employing two WMR, i.e. W2P30 and W3B62, respectively (cf. Figure 1 and sections of WP5.7). In accordance with the preferred importance of BLASTN derived items (see section WP5.6), the paths X2 →N→ X2 and X3→N→X3 were not comprised in our investigation. The possible usage of a similar strategy in cases of searches defined by WF19 and WF20 is not necessary due to the lower extent of the results.

Additional paths concerned BLASTN searches differing only in inputted queries MNSQ1 and MNSQ2:

N1 → N2 →N1, (WF38)

N2 → N1 →N2, (WF39)

where N1, N2 denote BLASTN with MNSQ1 and MNSQ2 respectively. Only sequence items selected by different MNSQ units of MNSQ1 and MNSQ2 were further processed in final selection steps.

WP2.3.4 Rules for co-localization and frequencies of MUSAS necessary for TCA. Subject segments co-localized with the corresponding dominant segments of superior score (**DSSS**) were included in the evaluation of MUSAS (see section WP5.7) frequency provided that: (i) the scores of DSSS and MUSAS were at least 25 and 22 bits, respectively; (ii) MUSAS overlapped at least 80% of DSSS length, if they were at least equally long as DSSS, or (iii) at least 80% of MUSAS similarities were present in DSSS-limited region, if they were shorter than DSSS.

Limits for frequencies of MUSAS and total scores were estimated based on a reference set containing the non-redundant version of records obtained from single-step BLASTN searches (limited with maximum scores of 30 bits; see section WP2.3.1) representing a sufficiently extended data set. This estimation was necessary due to too large records prevented to made direct evaluation, and also due to the assumed overestimation (yielding too strict limits) caused by a cumulative combined usage of parameters (frequency, bit score limits, presence in two searches). The assumed overestimation was in agreement with the number of TCA-derived MRNS lower than the number of the MRNS obtained by reference BLASTN searches. This means that even the employed enlargement of score limit by five bits in the reference set still did not appear to be able to fully prevent the assumed overestimation of frequencies (cf. WP2.3.2).

In more detail, the reference BLASTN derived data necessary for the limit estimation were processed using t-test [10] (for similar procedures see also [6]). This enabled us to determine minimum significantly different values (**MSDV**; $p < 0.05$). Data obtained with MNSQ1 as QS determined MSDV of six and five

for frequencies of MUSAS and the corresponding MSDV for total-score values 172.2 or 143.1 bits in searches with ER comprising (i) cancer-related or (ii) phosphorylation-related molecules, respectively. The related MSDV concerning MNSQ2 determined a limiting unique frequency of eight, whereas the corresponding total-score related MSDV were 226.4 and 254.0 bits. In accordance with (i) the enumerated MSDV, (ii) the current scale of bit-score-based BLAST classification of chain similarities (present in headings of current records of BLAST searches) and (iii) rounding-off-tendencies, we estimated **MUSAS frequency higher than five** and **total-score of at least 200 bit**s as limits for TCA.

WP2.4 Final evaluation and re-evaluation of terminally displayed MPL segments

WP2.4.1 Steps of final MPL selection and determination of conserved domain context. Separate sets of MRNS were obtained from the BLAST records when removing inner sequence redundancies in each record in **the first step** (cf. Figure 1 and Figure 1B in SF3). The subsequent comparisons using BLASTX version of program Align or current BLASTX enabled us to assess the protein segments containing aa encoded by at least a single nucleotide of MRNS, i.e. **MEPS**, in **the second step** and the sequences enveloping MEPS (**SEM**). SEM were larger than the included MEPS and contained at least 30 aa, if the both corresponding extensions of MEPS in N- and C-terminal directions were possible. This means that edges of at least ten aa extended MEPS in the left and right direction in the favorable cases.

Both database searches for empirically proved MEPS-related PPS and alternative prediction of such PPS occurred in **the third step** (for program tools see section 2.1). Predictions of phosphorylated aa present in MEPS were then limited by minimum scores of 0.800 in the two online evaluations. Stricter conditions for such selection concerned PPS on boundary lines of MEPS and their extensions. More precisely both predictions of phosphorylated aa occurring in the nearest, the second nearest and the third nearest neighbor aa to MEPS were limited by minimum score values 0.900, 0.950 and 0.950, respectively. The given evaluation of the three nearest neighbors followed from: (i) central position of phosphorylated site in the predicted nonapeptide PPS (determined by both programs) suggesting the importance of four aa neighboring the phosphorylated aa, and (ii) lower MRNS linkage to edge MEPS aa, which were only partially encoded by edge nucleotides MRNS and appeared thus to be critical for validity of overlaps including only unique forth aa. If any of all MEPS aa or any of aa present in the distance of three aa positions from MEPS was empirically confirmed as phosphorylated, then it was capable of being processed in further search steps.

In the **fourth step**, called here **feedback comparison**, we revised the positional agreement between PPS-related MRNS and MSEP segment of initial MSA-record. Align mediated TBLASTN similarities between MSA-record-derived MPQ3ab (MPQ containing MSEP chains and both LE and σ–consensi) and topical MRNS-related MNSQ segments (segments restricted by MRNS similarity of PPS relationship) limited by **minimum score 10 bits** proved required minimum (one aa) overlaps with CSB1. Additional combined usage of Align-mediated BLAST similarities then tested an **alternative requirement** to the preceding TBLASTN, i.e. **at least 50% extension of overlaps with CDR1** (cf. sections 3.4 and 4.2). In more detail, 50% overlaps of CDR1 were verified using three steps comprising (i) search for +/+ oriented sequence TBLASTN similarities between the whole MNSQ units related to the evaluated MRNS and

MPQ3ab, (ii) generation of alternative TBLASTX-derived identities of the whole MNSQ unit and its MRNS-related segment and (iii) restriction of in frame coherent CDR1-related segments in the two preceding searches.

In the **fifth step**, we observed the content of HM in the selected MRNS using Fuzznuc program (motifs WRCH and TCW and certain WRCH pairs, i.e. **W-pairs**, RGYWRCY, WRCWRCH, WRCHNWRCH, WRCHN(9)WRCH; cf. [11-15]). The number limits of **HM located at critical position with respect to possible aa alteration** (**HM***) considerably differed in final selection (one HM* was always necessary) and the later retrospective reselection (four HM* were necessary in case of failing in other reselection alternatives).

The **sixth step** represented an assembly of partial yet independent sequence records to a single united one. Since the **MPL-related items** were rearranged according to feedback-comparison-derived MSA-record-related mean positions (m(S); cf. Figure 5), it was simple to find neighbor items with the same SEM sequence but associated with different search paths. These items were then united to form a single item record non-redundantly summarizing all employed initial strategies.

Though the presence of **conserved domain context** was not necessary for the selection of the final set, it still represented a useful parameter for (i) further considerations or (ii) as an alternative condition of retrospective reselection if conserved Ig domain context was found (see Figure 5). To determine the conserved domain context, we preferred first of all regular conserved domain similarities, whose sequences contained a conserved domain segment identical with MPL-related MEPS. If such identity was not found, the co-localized conserved domain attaining maximum score and minimum Expect was presumed to represent the conserved domain context of evaluated MPL.

WP2.4.2 Origin of seven alternative criteria determining terminal retrospective reselection. All the **points of retrospective reselection** are briefly described in the fifth part of Figure 1B (cf. also the preceding section including shifted numbers of points with respect to Figure 1B located in SF3). For the facts necessary for derivation of rules mentioned in points (i) and (iii) restricting retrospective reselection (displayed in Figure 1B) see sections 3.3 and Figure 6. The requirement of more than two parallel strategies described in reselection point (ii) corresponds to the minimum discrete number level in which less than 5% of all items was found. Since the maximum frequency of individual WRCH or TCW occurrence in sole MPL was four, the same number determined the limit for overall HM occurrence (cf. reselection point (iv)). For specifically diversified limits restricting statistically deviated frequencies of repeatedly collocating similarities (obtained in different BLASTN searches limited by 30 bits), i.e. limits necessary for point (v), see section WP2.3.4. The alternatively required Ig domain context (point vi) appeared to be important with respect to the investigated structural relationship to AR. Except for a bit score limit for any recorded co-localizing similarity (10 bits), the reselection point based on feedback similarities (reselection point vii) kept the rules for TCA (see sections WP2.3.4, WP5.7 and third part of Figure 1B). The frequency limit of at least six feedback similarities then represented also the presence of more than half of the maximum number of such similarities, when recording only a single of two consensus relationships as consensus representative.

WP3 Comments to Figures

WP3.1 Time schedule of the main steps described in Figure 1. The initial evaluation of MSA-record and selection of CSB1 and CSB2 occurred in July and August 2013. The searches necessary for FFAS-scan slipped from August to November 2013. The searches determining MNSQ1 and MNSQ2 were performed from September 2013 to January 2014, when both MNSQ were assembled. MNSQ searches for MRNS were finished in July 2014, but the corresponding revising searches continued until September 2014. On-line translation was terminated in August 2014. Last selection of MPL and all steps described in Figure 1 were stopped in November and December 2014, respectively.

WP3.2 Combined statistical evaluation of graphs present in Figure 6. Maximum score products, representing each predicted phosphorylation site, were inputted to our model evaluation only if the predicted phosphorylation site was immediately encoded by MPL (cf. Figures 1 and 6 and section WP2.4.1). To evaluate regional differences within the distribution of score products displayed in Figure 6, we restricted pairs of exactly comparable regions in several steps (i.e. bottom **B(i,j)** and upper **U(j,k)**). During this "skeptical" regional restriction, we considered (i) local densities, (ii) reasons of no or very low occurrence of score products in the assumed range of extremely low product values and (iii) we regarded also critical individual occurrences or absences of score products in the limiting edge cases. The observed distribution was evaluated with respect to **x-intervals**, x-intervals - the intervals of score products with extent of 0.01 (i) quantifying axis x composing histograms in Figure 6 (cf. sections WP3.2 and WP3.3) and (ii) associated with the localizing values denoted here as x-values.

The differences between pairs of score values determined by NetPhos 2.0 and KinasePhos 2.0 (together with disregarded analyzed non-random effects) were assumed as the reason for low-density occurrence in the left part of the corresponding authentic histogram (the first histogram of Figure 6) in our model. In agreement with this fundamental assumption, we look for the most extended (critical) **bottom regions B(i,j)** with higher x-values, than those in the **region R** restricted based on random variation of the score differences under condition $p < 0.05$. The bottom x-interval of B(i,j) was denoted here x-value **b(w,a)** or **b(i=1)**. An additional attention concerned x-intervals with positive occurrence of score products in the region including x-intervals characterized by x-values **b(i>1) < b(i=1)**. To determine b(w,a) we first of all enumerated minimum significantly deviated differences between scores q(w,a,+/-) statistically. In fact, four values q(w,a) were derived based on t-test evaluation (cf. [10]):

$$q(w,a,+) = \text{mean}(\text{Abs}(y_i - z_i)) + t(1-w/a, df) \times d, \quad \text{(WF40)}$$

$$q(w,a,-) = \text{mean}(\text{Abs}(y_i - z_i)) - t(1-w/a, df) \times d, \quad \text{(WF41)}$$

where two possible a values (i.e. one and two) denote usage of one-side or two-side t-test, respectively; mean() is a mean value; y_i, z_i are maximum score values determined by NetPhos 2.0 and KinasePhos2.0 for i-th phosphorylation site, respectively; t(1-w/a, df) denotes the quantil of the employed t-test for $p < w$; df is degree of freedom (more precisely, $df = r - 1 = 39$ in case of r representing the number of all enumerated pairs of scores); d is standard deviation (known as σ_{n-1}).

The usage of both one- and two-side t-tests determined two negative values q(w,a,-) in our evaluation. This double-negativity (together with our skeptical considerations about the observed distribution) selected only left side of Figure 6 (bottom with respect to x-values) as a critically empty

region and justified thus the usage of one-side t-test. In agreement with this regional allocation, minimum selected scores Min(s) = 0.800 and q(w, a=1,+) (obtained in one-side test) determined precise position **b*(w)** within bottom x-interval position b(i=1) = 0.74 of B(i,j):

b*(w) = Min(s) x (Min(s) + q (w,1,+)) = 0.732798562. (WF42)

Nevertheless in accordance with b(w) value, the x-interval of b(i=1) = 0.74 is partially empty, which led us here to use secondary linear combination comprising results following b(i = 1) = 0.74 and b(i = 1) = 0.75 (cf. comments to two-step linear combination under formula WF53). Only a single x-interval of lower x-value than b(i=1) = 0.74 recorded occurrence of score product and determined thus the second possible bottom edge of B (i,j), i.e. b(i =2) = 0.72 (see the first histogram in Figure 6).

In the middle part of the first histogram of Figure 6, we disregarded x-interval 0.87 (without any score product value) as neutral zone between the dominant peak and other values. Consequently, we derived of edge neighboring x-intervals of non-overlapping compared regions of Figure 6, i.e. bottom B(i,j=1) and upper U(j=1,k) characterized by middle positioned values **m(j=1) = 0.86** and **n(j=1) = 0.88**, respectively.

Two types of possible errors followed from (i) limited accuracy of the determined scores causing differences in sorting the score products to sharply restricted of x-intervals and (ii) certain distribution of differences between scores. To minimize such errors, we constructed the two smoothened histograms (the second and the third pictures of Figure 6) simply estimating possible more robust distributions with respect to the considered errors (Figure 6 and section WP3.3). Both these histograms suggested less validity of x-intervals 0.86 and 0.88 (neighboring neutral zone of upper and bottom regions) indicated by overlaps of the constructed peaks. This determined second pair of the corresponding middle values **m(j=2) = 0.85** and **n(j=2) = 0.89** restricting regions B(i,2) and U(j=2,k), respectively. In accordance with (i) the regional specific densities of product occurrence and skeptic approximative considerations, (ii) the comments to smoothened histograms, (iii) negative values of q(w,a,-) in the both tested cases (none u(w,a)), and (iv) absence of score products in unique x-interval (single possible u(i>1) = 100 was found before as u(i=1)), the **regions of upper x-intervals of U(j,k)** contained only unique critical **u**pper x-interval of position **u = 1.00**, i.e. **U(j,k) = U(j)**.

The main statistical evaluation included enumeration of (i) odds ratios and (ii) significance levels of the differences between the observed regional distribution of score products. Both values were obtained based on **comparison with model constant distribution** using by one-side **Fisher´s exact test** [16] and **chi-square- (χ^2)** mediated evaluation [17] (for comments see section 3.3 and Figure 6). More precisely, we looked for w(i,j) values ($p < w$) of low difference from adequate p(i,j)-values simultaneously requiring strong and significant linkages, i.e. odds ratios $OR(i,j) \geq 2.0$ and $p < 0.05$. To simplify our descriptions in the following text, we used the characters **i, j present in parentheses** (substituting indexes) only in the cases of initial declarations or definitions and/or if new indexes appeared. In other cases, only empty parentheses were recorded, e.g. α() was used instead of α(i,j).

First of all, we assembled the corresponding 2x2 tables necessary for odds ratio evaluations independent of further changes in bottom table elements (because the following multiplication of both bottom elements with the same value yields the same OR but different p(i,j) or w*(i,j)):

α(i,j)	β(i,j)
γ(j)	δ(i,j)

where α(i,j) and β(i,j) are the numbers of score products in regions U(j) and B(i,j), respectively. The numbers of x-intervals in given regions were enumerated using formulas with the values b(i), m(j), n(j) and u mentioned above:

$$\gamma(j) = 100 \times \{u - n(j)\} + 1, \quad \text{(WF43)}$$

$$\delta(i,j) = 100 \times \{m(j) - b(i)\} + 1, \quad \text{(WF44)}$$

whereas the odds ratio OR(i,j) needed formula:

$$OR(i,j) = (\alpha() \times \delta())/(\gamma() \times \beta()). \quad \text{(WF45)}$$

Significance levels were calculated coming from the recorded overall numbers of evaluated score products (η(i,j)) and x-intervals (θ(i,j)):

$$\eta(i,j) = \alpha() + \beta(), \quad \text{(WF46)}$$

$$\theta(i,j) = \gamma() + \delta(). \quad \text{(WF47)}$$

In the initial step of our statistical evaluations, we enumerated four density aliquots, i.e. (i) tightly approximated integer numbers (**ι(i,j,m)** or **κ(i,j,m**)) and (ii) precise values (**λ(i,j)** and **μ(i,j)**), as necessary quantities representing compared values of constant distribution:

$$\iota(i,j,m=1) = int(\eta() \times (\gamma()/\theta())\} + 1 = int(\lambda(i,j)) + 1, \quad \text{(WF48)}$$

$$\iota(i,j,m=2) = int(\eta() \times (\gamma()/\theta())\} = int(\lambda(i,j)), \quad \text{(WF49)}$$

$$\kappa(i,j,m) = \eta() - \iota(), \quad \text{(WF50)}$$

$$\mu(i,j) = \eta() - \lambda(). \quad \text{(WF51)}$$

The approximated values then enabled corresponding web-page-mediated calculation of w*(i,j,1) and w*(i,j,2) determining validity intervals (w*(i,j,1) < p(i,j) < w*(i,j,2) using Fisher´s exact test:

α()	β()
ι(i,j,m)	κ(i,j,m)

or chi square enumeration

α()	β()
λ()	μ()

based on formula [17]:

$$\chi^2(i,j) = \{(\alpha + \beta + \lambda + \mu) \times (\alpha \times \mu - \beta \times \lambda)^2\}/\{(\alpha + \beta) \times (\lambda + \mu) \times (\alpha + \lambda) \times (\beta + \mu)\}, \quad \text{(WF52)}$$

To better approximate accessible results of Fisher´s exact test, we used **linear combination**. This possibility followed from convex-linear graphs of discrete **Fisher's test derived functions** p*(x), which values were enumerated after location of positive values a, b, c and d and different integer x-values (not higher than c) to the corresponding active web page 2x2 tables:

a	b
c-x	d+x

In accordance with convex profiles of p*(x), actual continuous **p(x) has to be always less strict than w(x**) enumerated by a linear combination of neighbor p*(x) values, when assuming continuous convex

relationship analogous to that of p*(x). Hence, in accordance with the given assumption, it holds: if w(x) = w than p(x) < w including topically critical w = 0.05.

The linear combination necessary for this purpose was unambiguously determined by formula based on differences between the overall and integer values (function int):

$$w(i,j) = \xi(i,j) \times w^*(i,j,1) + (1 - \xi()) \times w^*(i,j,2), \qquad \text{(WF53)}$$

where $\xi() = \mu() - \text{int}(\mu()) = 1 - (\lambda() - \text{int}(\lambda()))$ and int returns integer part of value.

Similar second step of linear combination was also necessary to include non-integer restriction of bottom value of B(i,j) with b*(w) (cf. formula WF42). The necessity of such two-step evaluation followed from recuired input of integer numbers to accessible web-page-mediated enumeration of Fisher test.

All OR(i,j) and w(i,j) indeed confirmed strong (OR ≥ 2) and significant (p < w < 0.05) associations of score products with upper regions compared to bottom regions in case of skeptical model regional restriction of the first histogram of Figure 6. The data are displayed in Table WPT1 (see below). Table WPT1 also confirms the limit score product limit 0.88 used as a posteriori limit in final reselection (cf. footnote c and Figure 1).

Table WPT1. Evaluation of the selected critical regions forming the first histogram in Figure 6

{i, j}	**B(i,j)**[a]	**U(j)**	**OR(i,j)**	**w(i,j)**[b]	**χ^2(i,j)**[b]
{1,1}	0.74*-0.86	0.88-1.00	2.49	0.0473	3.707
{1,2}[c]	0.74*-0.85	0.89-1.00	2.64	0.0420	3.925
{2,1}	0.72-0.86	0.88-1.00	2.69	0.0284	4.568
{2,2}	0.72-0.85	0.89-1.00	2.86	0.0247	4.855

[a]Extents of the selected intervals. * - precise b*(w) value (cf. formula WF42) restricted in fact special bottom value of enumerated intervals B(1,j), i.e. two step linear combination (mentioned under formula WF53) was used to determine w(i,j).
[b]The values determining validity. For enumeration of chi-square χ^2(i,j) and Fisher-test-derived w(i,j) see formulas WF52 and WF53, respectively. Both values represent in fact approximated numbers, but this approximation is grosser in case of χ^2(i,j) evaluation. Green and yellow labels: Though the chi-square is slightly lower than the necessary chi-square minimum enumerated valid value 3.841 (p < 0.05), the more precise value w(i,j) enumerated based on Fisher test still confirms the same validity.
[c]The interval of x-values 0.89-1.00 is restricted by the bottom value 0.88 equal to the proposed limit in section 3.3 , i.e. statistical parameters present in this row confirm and correspond to the limit employed in the paper.

WP3.3 Smoothened histograms based on data-consistent Gaussian kernel measures (displayed in last picture of Figure 6). To estimate more smooth histograms than original one (cf. the first picture of Figure 6) still adequate to represented data, we considered existing actual uncertainty ("diffusion") in determination of score-product values. This led us to use Gaussian kernel measures derived based on statistical analysis of processed (smoothened) data and performed in two-step process. In the first step of this process, we determined standard deviation value ($\sigma_{n-1}(1+2)$) evaluating differences between scores s1 and s2 as well as recorded score products s1*s2. In the second step, we utilize the corresponding

enumerated $\sigma_{n-1}(1+2)$ to access σ-value as necessary parameter specifying employed term of normal distribution function before we enumerated kernel measures.

The necessary two types of initial standard deviations $\sigma_{n-1}(i=1,2)$ were enumerated according current formula:

$$\sigma_{n-1}(i) = \{1/(N(i)-1)\} \times \Sigma_{n = 1 \text{ to } N(i)} (\Delta x_n(i))^2 \qquad \text{(WF54)}$$

where N(i) is number of evaluated differences from mean value in i-th evaluated set (N(1) = N(2) = 40 in our case); $\Delta x_n(i)$ denotes two sets of values, i.e. $\Delta x_n(1)$ and $\Delta x_n(2)$ defined by different formulas:

$$\Delta x_n(1) = s1_n\text{^}2 - s1_n \times s2_n, \qquad \text{(WF55)}$$

$$\Delta x_n(2) = s2_n\text{^}2 - s1_n \times s2_n. \qquad \text{(WF56)}$$

In fact, $\Delta x_n(i)$ values determined immediately $\sigma_{n-1}(1)$, $\sigma_{n-1}(2)$ as described in formula (WF54). Both enumerated $\sigma_{n-1}(i)$ then enabled us to count terminal resultant joined standard deviation value $\sigma_{n-1}(1+2)$, mentioned above, based on current formula [10]:

$$\sigma_{n-1}(1+2) = \{[(N(1)-1) \times \sigma_{n-1}(1)\text{^}2 + (N(2)-1) \times \sigma_{n-1}(1)\text{^}2]/(N(1)+N(2) - 2)\}\text{^}0.5. \qquad \text{(WF57)}$$

In the second step of our evaluation process, the following term composing general normal (Gauss) distribution was applied to enumerate values of kernel functions:

$$f(x,\sigma) = \exp\{(x-z)\text{^}2/[(-2) \times \sigma\text{^}2)]\}, \qquad \text{(WF58)}$$

where σ value was specified with help of the subsequent formula representing in fact transformation following from current interpretation of σ as inflex-point-related extent of Gauss curve [18]:

$$\sigma = (1/\tau) \times \sigma_{n-1}(1+2), \qquad \text{(WF59)}$$

where τ is extent of the evaluated x-intervals

The corresponding symmetrical kernel functions κ(k,ε,σ) were determined as "normalized values" of integrals:

$$\kappa(k,\varepsilon,\sigma^*) = \kappa(-k,\varepsilon,\sigma^*) = \{1/S(k,\varepsilon,\sigma^*)\} \times \int_{x = k - 0.5 \text{ to } k + 0.5} f(x,\sigma^*)\, dx = \{1/S(k,\varepsilon,\sigma^*)\} \times F(k,\sigma^*), \qquad \text{(WF60)}$$

where $F(k,\sigma^*)$ is enumerated value of integral; k denotes k-th neighbor x-interval counted from central representative x-interval (indicated by k = 0) when keeping the same (one-side) direction; ε is topical maximum of k-value; z = 0; σ* indicates terminally fixed σ-value; $S(k,\varepsilon,\sigma^*)$ represents sum of all ε-related $F(k,\sigma^*)$, i.e.:

$$S(k,\varepsilon,\sigma^*) = \Sigma_{k = -\varepsilon \text{ to } \varepsilon} F(k,\sigma^*). \qquad \text{(WP61)}$$

The repeating process of $\kappa(k,\varepsilon,\sigma^*)$ enumerations was started with ε = 0 (for ε = 0 holds $S(k,\varepsilon,\sigma^*) = 1$) and continued with the steps ε +1 → ε. This process was finished, if $\kappa(k= \varepsilon,\varepsilon,\sigma^*) < 0.005$ (indeed 5x10^-3), selecting thus last but one inputted value ε as sufficiently limiting value **ε***. The determined values ε* and $\kappa(k,\varepsilon^*,\sigma^*)$ enabled us to calculate resultant kernel measures ($\boldsymbol{\mu}_j$):

$$\mu_j = \Sigma_{k = -\varepsilon^* \text{ to } \varepsilon^*} \kappa(k,\varepsilon^*,\sigma^*) \times g_{j+k}, \qquad \text{(WF62)}$$

where g_{j+k} denotes number of products in (j+k)-th x-interval in unmodified distribution described in the first picture of Figure 6.

Based on empirical statistics, we determined here $\sigma_{n-1}(1+2)$= 8.287683 x 10^-3. Using this value in the Gaussian term (mentioned above and concerning x-intervals with extend τ = 0.01), we restricted ε* = 2 and then enumerated the corresponding statistically consistent measures forming smoothened histogram

of kernel measures μ_j (**"kernel histogram"**; see bottom picture of Figure 6). This histogram was important for restriction of dominant reselection limit for score product. Similarly to gliding mean frequency value at position 1.00 (cf. the second picture of Figure 6), the frequencies neighbor to right edge of the "kernel histogram" at positions 1.01 and 1.02 necessary for the enumeration of kernel measures at positions 0.99 or 1.00 were logically approximated with zero. Hence higher score values than one did not exist in given case. On the other hand, the frequencies at positions 0.69 and 0.68 necessary for enumeration of left edge values of the "kernel histogram" at positions 0.70 or 0.71, were actually equal to zero as well as all other possible lower x-values (cf. section WP3.2).

WP4 Comments to figure forms of tables

*WP4.1 An approach to significance of sequence block columns (**SBC**) determined in Figure 2.* Current BLAST enumeration determines the significance level of n-th SBC (p_n) :

$$p_n = 1 - \exp(-E_n), \qquad \text{(WF63)}$$

where E_n is Expect value in n-th SBC. Since single SBC aa identity is always enumerated, when optimizing non-invasive σ-consensus aa (cf. formulas WF1 and WF6 in section WP2.1), the following conditioned Expect value was used here to approximate column validity:

$$E_n = \{K \times \varphi_n(d,g,h) \times \exp[-\lambda \times (C_n^* - \gamma_n(g)]\}/\{K \times \varphi_n(1,0,h) \times \exp(-\lambda \times s_n[x_i,x_i])\}$$

$$= \eta_n(d,g,h) \times \exp[-\lambda \times \{C_n^* - (\gamma_n(g) + s_n[x_i,x_i])\}], \qquad \text{(WF64)}$$

For explanation of quantities and constants see sections of 2.2 and section WP2.1.

The corresponding restriction of valid columns was performed in accordance with significance level limits w (i.e. w determining p < w) for (i) profile-related PSI-BLAST evaluation (w = 0.005; indeed 5x10^-3; cf. sections WP2.2.2 and WP2.3.2) and (ii) the proposed SBC-related quasi-pattern (SQP) aa (w = 10^-6). Adding separately each of SQP aa to co-localized pattern aa, we can sometimes (i) improve the sensitivity of PHI BLAST supplementing new items, when keeping the same Expect limit, or (ii) determine different structural subsets of the compared molecules. In accordance with Figures 2 and 3 and the preceding text weak but existing aa similarities in SBC appear to be better classified and evaluated when using length equivalents, whereas statistical evaluation seems to be more valid and precise in cases of high aa similarities within SBC.

WP4.2 Subsets of the selected segments with distinct superior properties. In fact three types of superior properties concern MPL/MEP displayed in Figure 5 and "MPL-like MRNS"/MEPP shown in Table WPT2 (described below). These three types are described separately in the following three sections. The first type concerns hypermutation motifs (section WP4.2.1). The second one deals with extremes and extreme situations in domains and subdomain segments (section WP4.2.2). The third one then refers to extremes observed during steps of MRNS selection (section WP4.2.3).

WP4.2.1 Occurrence of hypermutation motifs. In accordance with nucleotide sequences of HM, all six alternative codons of serines (and consequently also with the obtained MPL-related data), the unique HM*-related phosphorylated aa of MEPS, i.e. serine, can be altered via TCW and WRCH distinctly present in transcribed and non-transcribed DNA strands, respectively. This means that, TCW present on transcribed DNA strand appear to be more perilous than WRCH with respect to knockouts of

phosphorylated serine due to proposed transcribed-DNA-strand-related additive effect of collision with transcription apparatus (cf. section 4.4).

HM critically located with respect to direct alteration of existing or predicted phosphorylated aa (i.e. **knockout of phosphorylation sites**) occurred in sixteen DNA segments displayed in Figure 5 and WPT2 (twelve MPL; unique reference MPL, i.e. NITR10; and three MPL-like, i.e. BRCA-1, GREB-1 and osteosarcoma oncogene; for Table WPT2 see below) and **concerned always serine**. Six such segments contained critically located WRCH (cf. human GREB1, titin, translation factor EIF4EBP1 and vascular endothelial growth factor receptor 2 or mouse Mia 3 and slowmo homolog 2) and ten such segments comprised similarly located but possibly more perilous (cf. the preceding paragraph) TCW (see human BRCA1, coiled-coil domain containing 88B, CS0DD006YL02 and tyrosine phosphatase PTPRZ1, or mouse glioma tumor suppressor candidate gene 1, osteosarcoma oncogene, potassium voltage-gated channel Kcnh7, tripartite motif-containing protein 30A-like and zinc finger protein 619 and/or reference NITR10 of *Miichthys miiuy* origin described in Figure 5 and WPT2).

Four of the selected MPL contained W-pairs in critical transcribed DNA strand at positions critical (with respect to at least a single WRCH) for aa alteration (cf. first and second paragraph of section 4.4). These MPL encoded two human proteins (forkhead box O1A present in rhabdomyosarcoma cells and zinc finger protein 687) and two mouse proteins (anaplastic lymphoma kinase and mouse zinc finger protein 619).

Six MPL contained maximum numbers of individual HM*. Four such WRCH motifs were present in each of three MPL encoding human translation initiation factor EIF4EBP (all WRCH formed two W-pairs), human BOC (without W-pair) and mouse zinc finger protein 619 (with single W-pair). Similarly, each of three MPL (composing tyrosine phosphatase PTPRZ1 or Rab11-binding protein of human origin and mouse glioma tumor suppressor candidate region gene 1) contained four critically located TCW.

WP4.2.2 Domain and subdomain relationships. Twenty-four of forty-four authentically selected human and mouse MEP composed segments achieving conserved domain similarity (conserved domain context) in Figure 5. Five of them attained context of conserved Ig domains. MPL encoding four of these five MEP co-localized during feedback comparison with the (i) MSA-record-related positions 15-22 of Figure 2 containing C-terminal segments of FR1 and (ii) neighbor segment of MSA record (positions 23-30) corresponding to the part of CDR1 region occurring only in light IgV chains, i.e. CDR1light (cf. Figure 2; at least 50% overlaps of CDR1light are indicated as CDR1L in Figure 5). However, the similarities of the given five MPL did not co-localized with more C-terminal segments of MSA-record (positions 31-35) limiting additional part of CDR1 of light chains and whole CDR1 of heavy IgV chains considered as common CDR1 segment, i.e. CDR1all (the corresponding sufficient overlaps are denoted here CDR1a; cf. Figures 2 and 5). Unexpectedly, considerable (more than 50% TBLASTN-related) overlaps of CDR1all were found during feedback comparison of MPL encoding adenylate cyclase 1 segment (domain Guanylate_cyc, i.e. pfam00211) and coatomer protein Copb1 (domain Coatomer_beta_C, i.e. pfam07718). Both these MPL encoded uniquely tyrosine-related PPS in accordance with previously described co-localization of tyrosine in both CDR1 and PKSIg (cf. [3]). In addition, considerable similarity overlaps of CDR1s (similarities mainly overlapping CDR1light but less

extensively also CDR1all segments; cf. Figure 2 and 5) were observed in cases of MPL encoding segments of EIF4EBP1 (involved in regulation of transcription; domain eIF_4EBP, i.e. pfam05456) and vav 1 oncogene (domain PH_Vav, i.e. cd01223) origin also without Ig domain relationship.

In addition to five Ig domain contexts of regularly restricted MPL, mentioned above, two types of domain contexts appeared two times in Figure 5 (SMC_prok_B, i.e. TIGR02168 and MAGE_N, i.e. pfam12440). Similarly to melanoma associated MAGE_N, three additional molecules were associated with melanoma, but only one achieved a conserved domain context with the domain PHA03247 (unambiguously titled).

Five MPL achieved TBLASTN-derived (direct but not feedback) similarities with the two MSEP-derived (LE- and σ-related) consensus sequences (of conserved Ig domains displayed in Figure 2) in the range of bit score values limited by the reference similarities of NITR molecules (NITR2 – 27.8 bits, NITR10 – 12.9 bits). Such bit scores were found in cases of human KDR (16. 8 bits), BOC (15.9 bits) or MAGEE1 (15.5 bits) and mouse adenylate cyclase 1 (17.2 bits) or tight junction protein ZO-3 (14.2 bits).

WP4.2.3 Extremes found during MRNS selection. More than two paths of selection strategies determined independently the same MPL in cases of two molecules. This concerns MPL of (i) human ret-oncogene (X12949.1) found by five paths and achieving also significantly increased frequency (**f**) of co-localized sequence similarities with different MNSQ1 units ($f = 8$; $p < 0.05$) and (ii) BOC appearing in three paths and representing a member of Ig superfamily as well as a unique molecule found by combined double-TBLASTX selection.

Two limits of the last alternative of retrospective reselection based on feedback comparison (cf. WP2.4.1) determined MPL of three molecules, i.e. human BOC (selected also by more than two alternative paths), human vascular endothelial growth factor receptor 2 and mouse glioma tumor suppressor candidate 1. The MPL of last molecule was unique achieving ten similarities with different MNSQ2 units when keeping the limit of 35 bits (significantly increased unit frequency $p < 0.01$; t-test; cf. section WP2.3.4).

WP4.3 MPL-like segments. Several functionally interesting molecules achieved parameters near the limits described above (see sections of WP2.4) and in Figures 1 and 1B. These molecules are shown in Table WPT2 which is in fact supplement to Figure 5.

Table WPT2. Interesting molecules near selection requirements (MPL-like segments)

No	Title of molecule[a,k] {numbers of SCTR}	Elementary information about MPL-like NS and PS[b] -sp[c] -str[c]	Clonal names (NS/PS)	Positions (NS/PS)	-mBS -FrS	Haa+ Waa[d]	Contexts of SEM[e] -SEM sequences and proposed HM-derived aa changes -local CD context(cdsBS,cdsE)	Existing/predicted PPS related to MPL[f,k] Pos +aa[f,g]	NetP[f]	KinP[f]	Spec[f,h]	Feedback comparison with the segments of initial MSA record[i,k] -mBS[b] -FrS[2b]	S,O m(S)[a]	- Aligned STS[j] - IgV positions[j,l]
46	Osteosarcoma oncogene {1}	Mu C2 K2	NM_010234.2 P01101.1	851-829 231-238	33.7 6	- + -	SLDLTGGLPEA**ST**P**e°s°e°EA**FTLPLLNDPE P **None**	12S **13T*P** 16S	0.962 - 0.988	0.976 0.881 - 0.948	ATM CK2 - ATM	14.2 1	7-11 4-11 9	AAB03680_igv FR1*
47	BRCA1 (breast cancer){5}	Hu C2	XM_006722041.1 XP_006722104.1	2074-2089 643-648	30.1 **14**	+ + -	PTDQLEWMVQLC**Ga*s°VVK**ELSSFTLGTGVH **BRCT smart00292 61.24 1.61e-11**	15S 21S	- -	0.895 0.936 0.891	ATM ATM CK1	14.6 1	8-11 6-11 9.5	AAB03680_igv FR1*
48	GREB1 (breast cancer) {0}	Hu C1	XM_006711908.1 XP_006711971.1	4929-4950 1503-1510 (C-term)	31.9 2	+ - -	SSPSPR**S**TTLS**SaNLEAsW**RACDYPS **None**	18S	0.926	0.906 0.805	ATM PKC	9.3 0	13-18 10-18 15.5	cd04981 (H_IgV) FR1C
49	CASC4 (cancer susceptibility) {0}	Hu C2	XM_006720381.1 XP_006720444.1	1391-1370 319-326	30.1 1	- + -	RIQTDILKQAT**KDr°VSd°FH**KLKQNDEEREL **None**	16S	0.984	0.822	ATM	23.5 8	15-21 14-22 18	cd00099 (IgV) + consensus FR1C
50	NITR2 (related to AR) {1,4,6,7}	Dl[m] DXS1	GQ385061.1 ACU27031.1	112-189 38-63	35.2 1+2	+ + #aW1 #aW9	SSLHFRSV**S**V**Ge°NVTLEcFYKDs#Ma#VKFY** **wYKqtPG**QKPQLMSKFN **IgV_TCR_alpha_like cd04983 + IgSF cl11960 both 83.84 3.10e-20**	9S	0.990	0.915 0.901	Aur ATM	46.4 10	17-41 16-41 29	cd04981(H_IgV) FR1C+CDR1s+FR2N

[a-j]For details see Figure 5.
[k]Gray background outside the column with “Contexts of SEM” indicates parameters insufficient with respect to regular MPL selection in the individual cases of MPL-like items.
[l]FR1* - framework region 1 without the required CSB1 overlaps (PIM: 5-14).
[m]Dl - Dicentrarchus labrax.

WP5 Additional details and explanations

WP5.1 FFAS-derived sequence scanning (FFAS-scan). FFAS comparison of whole CSB1 segments overlapping MSEP (see Figure 2 and section 2.1) with structures present in the databases PDB0113, SCOPE75 and PfamA26U selected only several sequences. The usage of minimum sequences (25 aa long chains) composing CSB1 segments and database PDB0113, improved the result in certain cases. It enabled us to determine a possible location of more conserved regions with respect to predicted fold relationships based on similarities with close paralogues of AR.

It was shown that FFAS-scan with the selected segments of cd00099, cd07706 and cd04980 yielded increased number of FFAS-scan-derived similarities with the paralogues of AR starting from –1 position of CSB1. On the other hand, such increase was not observed from the starting position 9 of CSB1, whereas the C-terminal segments inside CSB1 have starting positions 14 or 15-th according to numbers of gaps present in given segment of MSA record. This local restriction was in agreement with location of **pm3** consensus described in section 3.1. In accordance with this section, we selected eleven items of repeated pm3-related occurrence. Except for novel immune-type receptors 10 and 11 absent in mammalian proteome, nine protein items were present in the investigated human and mouse proteomes; all of them exhibited Ig superfamily relationship in conserved domain searches. Eight items exhibited TBLASTN sequence similarities to FFAS derived sequences (important with respect to paths of comparisons employed in this paper), i.e. programmed cell death protein 1, hepatitis A virus receptor homolog (both of superior IgV relationships), CD8 alpha chain, natural cytotoxicity triggering receptor 3, three isoforms of CD300, trem-like transcript 1. On the other hand, though V-set containing protein 1 (VCP1) was selected by FFAS scan, only other isoforms of VCP were found in these TBLASTN records comparisons. In spite of these promising relationships, only NITR10 mRNA contained terminally selected MPL sequences, whereas additionally selected NITR2 was near to the limits.

*WP5.2 A short overview concerning length equivalents (**LE**).* Sequence block columns (**SBC**) in record of multiple sequence alignments (**MSA**) represent the most frequent subjects of LE evaluation and a unique type of such subjects undergoing LE evaluation in the associated paper (for other types of LE evaluations considering neighbor columns or sequence block segments see our paper [3]). LE in fact represents the values of SBC probability or conditioned Expects by means of non-integer height of reference SBC containing only identical aa. LE thus determines the natural “corpuscular-related” hierarchy of model column identities determining LE, which can form a fuzzy-related system (cf. Figure 3 and see below). LE can be enumerated based on (i) formerly suggested binomial evaluation [3,19], (ii) substitution-score-related evaluation (see sections 2.2.1, 2.2.2 and WP2.1) and (iii) possibly also when using Dirichlet distribution (cf. [20]). Fuzzy-related LE-derived evaluation represents in fact (cf. also 2.2.2 and Figure 3) an objective natural classification of SBC similarities indicating and distinguishing among others sequence block segments with cumulating low but still important validity. This means that LE values can make more easy process of CSB restriction in MSA record with broad repertoire of SBC similarities (see sections of 2.2).

Some fuzzy-related limits derived based on length equivalents (**LE**) were described in the preceding papers [2,4,6] within the frames of the proposed approach called Evaluation of sequence similarities using

Length Equivalent Measures as a System (ELEMS). In fact three types of limits appeared to be important for ELEMS: i.e. (i) positively and negatively restricted deterministic (i.e. strictly restricted) limits, (ii) fuzzy related limits and (iii) empirically derived ones.

Negatively restricted deterministic limits fully discriminated the presence of subdominant categories. For instance minimum "fuzzy-related" similarity in SBC was defined in a binomial version of ELEMS as non-random quasi-similarity/similarity (for limit 1.5 see [3]). In accordance with the fuzzy-related terms proposed in this paper, LE value of 1.5 were approximately classified as a value present in the upper boundary line of major noise effects (see Figure 3). In contrast to the preceding limits, **positively restricted deterministic limits** were equal to LE value corresponding to maximum measure, e.g. limits deterministically restricting at least double sequence similarity (LE = 2) and strict cohesivity in SBC (LE = 3; cf. [4]).

Edges of intervals composing current fuzzy systems represent the most frequently used **fuzzy-related limits** (**FRL**; cf. Figure 3) though other strategies of defuzzification yielding FRL also do exist [21]. In addition to limits following from fuzzy formalization, empirical analysis accepts delay effects during the modeled dynamic events and regards statistics as well as evaluation of noise or errors in the corresponding data [21]. These **empirical limits** appear to be interesting mainly for an evaluation of phylogenically more distant structural relationships.

WP5.3 LE-derived fuzzy system. In this paper, we demonstrate a substitution matrix derived fuzzy-related system (**SMFS**; Figure 3) as a possible alternative subsystem of ELEMS. This system is more diversified than the current fuzzy systems, which frequently contain only seven intervals. In fact, SMFS includes all positive and negative single-character values, i.e. nineteen different values. Nine positive categories comprise three similarity-related categories (similar, cohesive and rigid), three of their quasi-forms, a single improved category (for a non-pairing odd upper value of nine) and two categories of low validity relationships (rampart of randomness and area of "promising" noise). In accordance with this classification, the description of intervals with negative F-values can be approximated using anti-categories to positive ones (e.g. at least anti-cohesive aa are dealt with the section 3.1 or area of "disagreeable" noise appear to be interesting). Given single-number description following from SMFS is suitable not only with respect to the possible short two-row record with positive and negative values under MSA record (see Figure 2) but also in case of further analysis of the MSA record forming a sequence block (cf. sections 2.2.2, 2.2.3, WP2.1 and WP2.1.2). The usage of quasi-forms in division and description of fuzzy-related intervals is considered as more adequate with respect to existing uncertainty in decisions concerning observed SBC similarities. This uncertainty appears to be important for: (i) buffering of possible errors following from single character values of substitution matrix scores (used also in the following BLAST searches for MRNS) and (ii) diminishing of possible random effects. The usage of modified enumeration of gap penalties in vertical direction represents an isotropic approximation keeping usual BLAST routines. As follows from the formulas presented in section 2.2.3, this vertical enumeration was not on the other hand included in the sequence block evaluation.

WP5.4 Protein queries determining nucleotide sequence queries MNSQ1 and MNSQ2 necessary for PPSIg-related database searches. Initial sequence searches leading to the formation of MNSQ1 (**ISS1**)

and MNSQ2 (**ISS2**) were performed with two and four types of individual sequence queries, respectively. Two common types of sequence queries were employed in both ISS1 and ISS2. This comprised (i) segments of typical sequence, i.e. STS-queries (cf. WP2.2.2), forming the first conserved segment of initial MSA record (CSB1) and (ii) unique CSB1-related consensus, i.e. consensus 1 (see above). All CSB1-related segments composed queries in ISS1, whereas only three CSB1-related segments of cd00099, smart00406 and igw (AAB03680) constituted ISS2-related QS selected in accordance with our preceding experience with conserved immunoglobulin domains [4]. In addition, four other sequence segments supplemented QS repertoire in ISS2, i.e. (i-iii) three specifically constructed hybrids (HQS) of STS-queries and consensus 1 [4] (cf. WP2.2.2), and (iv) empirically selected sequence segment called cls and achieving superior ternary similarities during ISS2 (see Figure 2 and Supplementary file2). All three types of multi-sequence-protein queries (**MPQ**) (i.e. different five MPQ) were necessary for revision procedures reassuming the results of both ISS1 and ISS2. These MPQ contained sixty character long units, each composed of the compared protein sequence and separating a spacer sequence with repeating X. More precisely the list of MPQ comprised MPQ1, MPQ2a, MPQ2b, MPQ3a and MPQ3b, described in sections 2.3, WP5.4 and WP5.5.

WP5.5 Some details concerning MPQ3 restriction. In accordance with the results of FFAS-scan described in section 3.1, we extended the PPSIg-containing sequence block segment: (i) to be equally long as CSB1 (see Figure 2), (ii) to be more shifted to the N-terminal SBC than CSB1, and (iii) to contain the whole sequence block segment related to pm-3 consensus in its C-terminus [3] (cf. section 3.1 and Figure 1B in SF3). In addition to all sequences of the restricted MSEP block segment, the corresponding LE- and σ-consensi segments constituted MPQ3a and MPQ3b, respectively. Both MPQ3a and MPQ3b were necessary for the selection of sequence segments forming units of both MNSQ (cf. sections 2.2, 2.3, WP5.4, WP5.5, supplementary file SF3 and Figure 2)

WP5.6 Reasons for the usage of single-step BLASTN selection and combined cumulative approach based on two currently but differently adjusted TBLASTX. Since the important hypermutation and insertion-deletion changes in compared immunoglobulin segments (forming MNSQ1 and MNSQ2) usually depend only on DNA sequences compared directly by BLASTN, single-step TBLASTX comparison appeared to be considerably less important for the corresponding research. In contrast, the combined cumulative approach based on collocation of sequence similarities in two types of TBLASTX (WMR: W2P30 or W3B62) was similar to rules for TBLASTN restricting the compared MNSQ (cf. section WP2.2.4) and also mostly yielding collocating similarities. Consequently, the given procedural similarity led us to include combined cumulative TBLASTX approach in the employed procedures.

WP5.7 Additional notes on TCA. TCA comprised three different restrictions: (i) frequency (more than five MUSAS) or related total-score based choices (a limit of 200 bits for total score) in accordance with section WP2.3.4, (ii) score limits always applied in both sequence searches (score limits of 25 bits and 22 bits concerned maximum segment similarities and MUSAS, respectively) and (iii) simultaneous occurrence of collocating segments (regularly similar in accordance with the preceding points (i) and (ii)) in pairs of two different types or two types of differently adjusted BLAST searches. In **initial searches** of TCA, total-score limits (total score values were still accessible via the starting part of the BLAST records)

approximated in fact the MUSAS frequency (for details see section WP2.3.4). Hence the overall records, necessary for MUSAS frequency evaluation, were too large and thus inaccessible independently of the communicating computer. Since TCA comprised two types of different searches (cf. Figure 1; for the corresponding paths of search combinations see WP2.3.3) the second different type of the searches (**complementary searches**) occurred without capacity problems, when using a limited list of clonal names extracted from the record of an initial search. The subsequent **reverse searches** in fact reproduced the initial search under conditions of favorable capacity. Hence by using repeatedly reduced lists of clonal names (obtained from a limited part of the complementary search record) we were able to simply perform evaluation of MUSAS frequencies.

WP5.8 Additional trends in bioinformatic investigation of MPL. The additional trends would concern: (i) similar search for MPL with phylogenically diversified MNSQ3 and MNSQ4 of TCR origin, (ii) mapping and statistics of differences in translation reading frames on a broader set of MPL, (iii) importance of secondary structures of nucleic acids (cf. [22]) and encoded peptides, W-pairs and CDR1-like segments found by feedback comparison (sections 3.4 and 4.4) with respect to the considered hypermutation effects in MPL, (iv) phylogenic analysis of MPL diversity (sections 3.3 and 3.4), (v) optimization and improvement of MPL selection (section 4.2), (vi) further specification of the proposed comparison between sequence blocks and sequences comparing given approach with Sequence logos, positional weighting and trends in evaluations of MSA records [23-26], (vii) reevaluation of HM*-related aa using future more complex and sophisticated databases of existing mutations (i.e. extended multi-species databases of actual mutation sites including regular selection of frequent or pathogenically associated mutational ones and their comparison with recorded hot spots), and (viii) extending combinatory repertoire of PPS prediction methods, to diminish indicated false negativities.

WP6. References to Supplementary File 1

[1] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

[2] Wilson P, Liu YJ, Banchereau J, Capra JD, Pascual V (1998) Amino acid insertions and deletions contribute to diversify the human Ig repertoire. Immunol Rev 162:143–151.

[3] Kubrycht J, Borecký J, Sigler K (2002) Sequence similarities of protein kinase peptide substrates and inhibitors: comparison of their primary structures with immunoglobulin repeats. Folia Microbiol 47:319–358.

[4] Kubrycht J, Sigler K, Ružička M, Souček P, Borecký J, Ježek P (2006) Ancient phylogenetic beginnings of immunoglobulin hypermutation. J Mol Evol 63:691–706.

[5] Kubrycht J, Novotná J (2014) Sequence-based prediction of linear autoepitopes involved in pathogenesis of IPAH and the corresponding organism sources of molecular mimicry. Int J Bioinf Res Appl 10:587–612.

[6] Kubrycht J, 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. J Theor Biol 255:8–15.

[7] Cooper GM (1995) Oncogenes, 2nd edition. Jones and Bartlett Publishers, London and Boston.

[8] Litman GW, Hawke NA, Yoder JA (2001) Novel immune-type receptor genes. Immunol Rev 181:250–259.

[9] Cannon JP, Haire RN, Magis AT, Eason DD, Winfrey KN, Hernandez Prada JA, Bailey KM, Jakoncic J, Litman GW, Ostrov DA (2008) A bony fish immunological receptor of the NITR multigene family mediates allogeneic recognition. Immunity 29:228–237.

[10] Zvárová J (2001) Biomedical statistics. I. Fundamentals of statistics for biomedical areas. Karolinum, Prague.
[11] Rogozin IB, Kolchanov NA (1992) Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochim Biophys Acta 1171:11–18.
[12] Dorner T, Foster SJ, Brezinschek HP, Lipsky PE (1998) Analysis of the targeting of the hypermutational machinery and the impact of subsequent selection on the distribution of nucleotide changes in VHDJH rearrangements. Immunol Rev 162:161–171.
[13] Beale RCL, Petersen-Mahrt SK, Watt IN, Harris RS, Rada C, Neuberger MS (2004) Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. J Mol Biol 337:585–596.
[14] Rogozin IB, Diaz M (2004) Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. J Immunol 172:3382–3384.
[15] Roberts SA, Sterling J, Thompson C, Harris S, Mav D., Shah R, Klimczak LJ, Kryukov GV, Malc, E, Mieczkowski PA, Resnick MA, Gordenin DA (2012) Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Mol Cell 46:424–435.
[16] Agresti A (1992) A survey of exact inference for contingency tables. Stat Sci 7:131–153.
[17] Lepš J (1996) Biostatistics. University of Southern Bohemia, České Budějovice.
[18] Kunderová P (1997) Introduction to theory of probability and mathematical statistics, University of Palacky, Olomouc.
[19] Kubrycht J, Borecký J, Souček P, 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 Microbiol 49:219–246.
[20] Altschul SF, Wootton JC, Zaslavsky E, Yu YK (2010) The construction and use of log-odds substitution scores for multiple sequence alignment. PLoS Comput Biol 6:e1000852.
[21] Jura P (2003) Fuzzy systems. Fundamentals of fuzzy logics for regulation and modeling. Vitium, Brno.
[22] Wright BE, Schmidt KH, Davis N, Hunt AT, Minnick MF (2008) II. Correlations between secondary structure stability and mutation frequency during somatic hypermutation. Mol Immunol 45:3600–3608.
[23] Schneider R, Sander C (1996) The HSSP database of protein structure-sequence alignments. Nucleic Acids Res 24:201–205.
[24] Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18:6097–6100.
[25] Dong E, Smith J, Heinze S, Alexander N, Meiler J (2008) BCL::Align-sequence alignment and fold recognition with a custom scoring function online. Gene 422:41–46.
[26] Thomsen MC, Nielsen M (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res 40:W281–W287.