Web page supplement to paper:

Sequence based prediction of linear autoepitopes involved in pathogenesis of IPAH and the corresponding organism sources of molecular mimicry

WP1. Some important features concerning our evaluation of DQPA occurrences in sequence database (i.e. SPECIES_VALENCE approach). SP_VAL is based on four important features of our evaluation. First of all, the consistent threshold of the selected DQPA represents an important parameter of specificity together with overall BLAST evaluation of DQPA similarities. Secondly, SP_VAL as propensity-related approach makes it possible to process less favorable DQPA identities evaluated as important in some cases of higher frequency. Thirdly, implemented empirical statistics of SDF and EDO values can be used even under otherwise hardly processable conditions. Fourthly, alternative models are included to better analyze different aspects of evaluated reality.

WP2 DQPA as potential autoepitope-related structures. Sequences of both F1SA and autoantigens represent in fact knowledge-based reservoirs of autoepitope structures. Consequently, dense short DQPA present simultaneously in both types of sequences form a concentrated set of potential autoepitopes (cf. 2.1). Since the set of IPAH related autoantigens is possibly still incomplete, we classify here preselected organisms (with DQPA related proteins) according to the presence of DQPA in IPAH related autoantigens or in a more extended set of autoantigens restricted by ScanProsite program and the proposed HMG relationship (Morse et al., 1996; cf. Table 3).

WP3 Nonapeptides form the longest DQPA-related sequences in our model. Nonapeptides interact with antibodies recognizing linear epitopes with high affinity (Königs et al., 2000; Afonin et al., 2001). Stereochemical models of

antibody interaction with nonapeptides then suggest complete overlap of antibody binding site by nonapeptides (Afonin et al., 2001).

WP4 Expansion to eight model situations. Different model situations in fact reflect considered variability of possible relationships, alternative explanations or existing events including current behavior of bacterial infection (cf. section 2.1 and Table 3). In addition, we test here the stability of achieved differences with respect to given variability, when using mean values and criteria mentioned in 3.3.

WP5 Some aspects of information bias. The set of organisms involved in respiratory diseases S1.2 is favorably optimized owing to i) more extensive sequencing in the area of medical microbiology and ii) frequencies of DQPA. Consequently, the application of the two different approaches of empirical statistics based on S1.2-related RSO set of hashes as reference set can usefully restrict limits with respect to contemporary status of sequencing, possible random relationships and some false positivities. This is why we speak here about organisms outside our selection instead of definitive decision.

WP6 References to reference set S1.2. All organisms of the reference S1.2 set (cf. also section 2.3) displayed here represent pathogens or commensals involved in acute bronchitis (MacKay et al., 1996), lung abscess (Mori et al., 1993; Hirshberg et al., 1999), acute pharyngitis (Bills et al., 1996), tuberculosis and pneumonias. The last group of diseases then includes **viral** pneumonias (Oda et al., 1994; Chien and Johnson, 2000), HIV-associated **opportunistic** pneumonias (Huang and Crothers, 2009; King et al., 2009), **bacterial** pneumonias (Marrie et al., 1998; Niederman et al., 2005; Yildiz and Dagonay, 2006), **fungal** pneumonias (Conces, 1999; Saubolle, 2000) or **differently**

(otherwise) **classified** pneumonias (Azevedo Sias et al., 2009; Niederman et al., 2005).

WP7 Formulas necessary for evaluation of DQPA set. The limit of valid sequence probability (P < MAX(P), where MAX is the maximum possible value under the condition p < w; w is 0.05 in our case) for local occurrences of **nonrandom dense aa sequence** can be derived using different procedures. In accordance with binomial statistics, we determined MAX(P) = 2.5063 x 10^-3 (cf. Lepš, 1996; Kubrycht et al., 2006; Kubrycht and Sigler, 2008). An alternative slightly higher probability limit (MAX(P) = 2,7625 x 10^-3) follows also from a case when both *a priori* and *a posteriori* Bayesian probabilities are limited by the same significance level p < 0.05 (i.e. p < w; w(P|w) = 0.05; w = 0.05). The corresponding formula is in fact equation in the case of unknown limiting P = MAX(P) (cf. Zimmer, 2004):

$$w(P|w) = P \times (1 - w)/(P \times (1 - w) + (1 - P) \times w), \quad \text{(WF1)}$$

Since local probability values of sequence identities (LPSI) are usually higher than the corresponding local random chain probabilities mentioned above, we can assume that the conventional limit used for pattern-related PSI-BLAST restriction p < 0.005 (Altschul et al., 1997) is sufficient to limit LPSI related Expects of DQPA identities. This Expects represent in fact minimum possible Expect of each query sequence ($\mathbf{E_{min,i}}$):

$$E_{min,i} = K \times H \times L_i\text{^}2 \times \exp(-\lambda \times S_i). \quad \text{(WF2)}$$

where K, H, λ are usual BLAST constants for ungapped BLASTP search (K = 0.2891920; H = 1.935981; λ = 0.294); L_i is length of i-th DQPA.

Significance levels (p < w) of $E_{min,i}$ easily follow from general BLAST-related formula (Altschul, 1990), which holds also for E = $E_{min,i}$:

$$w = 1- \exp(-E), \quad \text{(WF3)}$$

If we assume that addition of sufficiently restricted DQPA-related chains to the sequence set cannot negatively influence Expect evaluation, then overall

Expect evaluation related to multiple DQPA occurrence in the database depends mainly on individual Expects of lowest values. Hence such Expects indicate the similarities of the highest validity. The contribution of other DQPA then increases with the number of their identities. Under such conditions, the corresponding relationship between individual Expect values ($\mathbf{E_i}$) and overall expect ($\mathbf{E_h}$) is related to the EDO evaluation mentioned above and can be modeled by harmonic evaluation:

$$E_h = N_w \text{ x } E_1 \text{ x } E_2 \text{ x } E_3 \text{ x } E_n/\{(E_2 \text{ x } E_3 \text{ x...x} E_n) + (E_1 \text{ x } E_3 \text{ x... x } E_n) + ... + (E_1 \text{ x } \text{ x } E_2 \text{ x } E_3 \text{ x... x } E_{n-1})\} = N_w/(1/E_1 + 1/E_2 + 1/E_3 + ...+ 1/E_n), \quad \text{(WF4)}$$

where N_w denotes the number of window shifts restricting here the length of linear epitopes and can be easily enumerated, i.e. $N_w = (\delta - (W - 1))$. In accordance with the usual linear epitope extent we employed the window of length $W = 9$ (cf. section WP3) to cover F1SA sequence of length $\delta = 123$, i.e. $N_w = 115$.

The modified formula for individual Expect enumeration differs from the original BLAST formula only in the existence of specific DQPA-related database length M_i (cf. also formula 2):

$$E_i = K \text{ x } H \text{ x } L_i \text{ x } M_i \text{ x } \exp(-\lambda \text{ x } S_i) \quad \text{(WF5)}$$

More precisely, M_i denotes effective database length of huAG corrected with respect to i-th DQPA length:

$$M_i = D - Q \text{ x } (L_i - 1) \quad \text{(WF6)}$$

where Q is the number of autoantigen molecules ($Q = 140$); D is full database length, i.e. $D = \Sigma_{j = 1 \text{ to } J} C_j = 95\,469$, where C_j are chain lengths of separate sequences.

Full database length of huAG, i.e. D, is the sum of all chain lengths of the sequence set comprising autoantigens linked to IPAH, autoantigens forming a pre-restricted explicit autoantigen subset in ScanProsite, and HMG proteins of not fully restricted linkage to IPAH and autoimmune diseases. HMG were restricted by BLASTP comparison with RPS-BLAST-derived consensus of

HMG superfamily-related conserved domain as a sequence query (conventionally limited by $p < 0.005$) and reevaluated for conserved domain HMG-related similarities.

Like other sequence similarities of length 5-15 aa length, 5-9 aa long DQPA identities can be evaluated by means of "BLASTP search for short near exact matches" (**SNEM BLASTP**) using PAM30 substitution matrix. SNEM BLASTP is restricted by limits two hundred times higher than usual BLASTP. Consequently, the modified significance level of the selected DQPA set ($p < w_{DS}$) can be approximated by the following formula:

$$w_{DS} = 1\text{-}\exp\{- E_h/200\} \qquad \text{(WF7)}$$

Provided that enumerated DQPA similarities are sufficiently robust (see section WP8), we can use current significance limit $p < w_{DS} \leq 0.05$. In such case, we speak here about the valid SNEM BLAST similarity based on $p < w_{DS}$.

Related evaluations of DQPA occurrences in the compared microorganism sequence sets were not performed. Hence the majority of Expects values was higher than all usual significance limits due to frequent cases of database lengths higher than 10^6. Moreover, additional problems moreover follow also from the incompleteness or redundancy of current databases. Such state thus in fact implicates usage of empirical statistics described in sections 2.7-2.9 as a possible starting point.

WP8. DQPA set is sufficiently robust to use $p < 0.05$. Since random binomial and Bayesian evaluations are highly related in the range of usual limiting values $w = 0.05$ (Kubrycht et al., 2006), it is possible to delimitate significance limits using combined evaluation. Based on binomial statistics, we can enumerate the validity of a set containing s structures (i.e. set of sample size s) of significance level restricted by $p < w[s]$ (Lepš, 1996):

$$w[s] = (r \times (1\text{-}r)/s)\text{^}0.5, \qquad \text{(WF8)}$$

where r is a random chain probability. Formula WF8 enables to derive simple recurrent enumeration:

w[s] = (1/s)^0.5 x w[1] (WF9)

Formula WF9 and w[1] value are then necessary for the enumeration confirming the validity of w[s] with respect to sample size and predetermined significance level w. The evaluation is in principle identical with the evaluation represented by formula WF1:

w(w[s]|w) = (1-w) x w[s]/{(1-w) x w[s]) + w x (1- w[s])} (WF10)

As mentioned in section WP7, the value w[1] = P = 0.005 sufficiently restricts the pattern-related minimum Expect values corresponding to the limit of PSI-BLAST searches and also our DQPA comparisons. In accordance with this fact, we used a more strict combined Bayesian/binomial approach defined by formula WF10 (cf. section WP7) and unfavorable maximum w[1] to delimitate a limit of s stricter than the true limit, but still sufficient for our decision. More precisely, we determined the minimum number of DQPA, whose Expect values are approximated by the enlarged values of unfavorable maximum w[1] = 0.005 and their occurrence is sufficient for a'posteriori value w(w[s]|w) $\leq$ 0.05 (defined by formula WF10 under the validity condition w = w_{DS}; cf. formula WPF7).

The enumerated critical sample size of the DQPA set was twelve in our case, which in fact corresponds to s $\leq$ 12 in case of less strict Expect evaluation. This means that our set with 33 initial DQPA is sufficiently robust with respect to significance limit p < 0.05 for Expect value.

WP9 Why we use quasiconsistent level. The usage of **quasi-consistent significance level p < 0.10** is sometimes proposed in statistical tables as maximal limiting value instead of usual limit 0.05. Similarly, "negative big" values in fuzzy logic are limited by interval range 8.33% but not 5% (Jura, 2003), whereas the binomially derived significant occurrence level of

specifically recognized tetranucleotides achieves $p < 0.0896$ (Kubrycht and Sigler, 2008).

WP10. Additional relationships of DQPA to metalloproteinases and interleukins. Sequence of metalloproteinase ADAMTS16 contains other sequence VPRPP (Q11) of human SOX13 origin achieving the second REF related c_{im}. Q11 constitutes PAB-dependent poly(A)-specific ribonuclease subunit pan2 of both the selected Aspergillus species. DQPA YPPPP (Q5) of human Annexin A11 origin contains frequently immunogenic tyrosine together with multiple flexible prolines (i.e. potential flexible tyrosine; cf. section 4.1). This DQPA achieves forth highest value of cim among all DQPA found in microorganisms. Q5 constitutes five different sequences of proteins synthetized by both selected pairs of *Mycobacterium* and *Aspergillus* genera and was also found in the sequence of human interleukin 34.

Metalloproteinase ADAMTS8 (found by using three different DQPA5) inhibits angiogenesis (Vázquez et al., 1999). Model ADAM15 deficient mouse (ADAM15 includes DQPA6 frequent PPPPXXPXP) shows reduced neovascularization compared with wild-type controls (Horiuchi et al., 2003). Interleukin 6, including *E.coli*-related DQPA5 PVPPG in its sequence, is a pro-inflamatory cytokine found in significantly higher amount in patients with severe IPAH (more than four times higher mean values) (Humbert et al., 1995).

WP11. Structural changes of proline - additional comments. Since the energy difference between trans- and cis-configurations of prolyl residues is very low in proteins, the corresponding **trans/cis-isomerization** can occur even spontaneously (Schmid, 1993; Chiang et al., 2009). In addition, this isomerization is also accelerated by prolyl-isomerases (Schmid, 1993; Wang and Etzkorn, 2006). The group of human proteins with prolyl isomerase activity includes cyclophilins, FKBPs, and parvulin, although some larger proteins also

contain prolyl isomerase domains. Cyclophilin B present in human plasma and some molecules involved in infectivity of lung pathogenic organisms, e.g. *C.neoformans* selected here (Table 3), *Streptococcus pneumoniae* and *Legionella pneumophila* (Cianciotto et al. 1989; Ren et al., 2005; Hermans et al., 2006) represent the possible prolyl-isomease sources in the extracellular space. This implicates the question of possible additive effects of bacterial prolyl-isomerases in the mechanism proposed here (e.g. in case of co-infecting pathogens). Chaperon activities of prolyl-isomerases were recently analyzed on structural level (Jakob et al., 2009). Hydroxylation of prolyls usually stabilizes fibrous poly-proline helix II. This means that the effects diminishing this **hydroxylation** can also participate in prolyl-related structural changes (Adzhubei and Sternberg, 1993; Adzhubei and Sternberg, 1994). Antibodies against proline rich epitopes have been screened in several diseases (De Keyser et al., 1992, Greene et al., 2010).

WP12. Some additional notes to the selected microbes and EDO values. *Pseudomonas aeruginosa* represents together with *Staphylococcus aureus* and *Acinetobacter baumannii* an organism of high frequency during hospital-acquired pneumonia (Beardsley et al., 2006) and is associated with high mortality rate during lung abscess (Hirshberg et al., 1999). Invasive aspergillosis is perhaps the most devastating of *Aspergillus*-related diseases, targeting severely immunocompromised patients (patients with AIDS and other hematological malignancies such as different leukemias, solid-organ and hematopoietic stem cell transplants, genetic immunodeficiencies such as chronic granulomatous, or on prolonged corticosteroid therapy; Dagenais et al. 2009). Noninvasive aspergillomas may form following repeated exposure to conidia and target preexisting lung cavities such as the healed lesions in tuberculosis patients (Dagenais et al. 2009). In multivariate analysis, the only significantly positive influencing factors on colonization by *Candida albicans* at the tongue

were tongue piercing (p = 0.034) and daily smoking of more than 10 cigarettes (P = 0.024; Zadik et al. 2010). The increased occurrence of Candida (predominantly *C. albicans*) in saliva and faeces of the psoriatic patients suggests at least adjuvant effects of this organism (Waldman et al., 2001). *Mycobacterium bovis* belongs to the group of highly related microorganisms (at boundary line of species and genus; perhaps in the range of sub-genus taxonomy) denoted as Mycobacterium tuberculosis complex and including also *M. tuberculosis*. This microorganism can also cause human tuberculosis and possibly also at least some of the autoimmune effects described in section 4.3 (O'Reilly and Daborn, 1995; Dubaniewicz et al., 2010; Homolka et al. 2010). Symbiotic bacteria *E. coli* causes community-acquired pneumonia in some cases (Marrie et al., 1998; cf. Table 3 and sections and 3.3 and 4.5).

In the case of cellular and combined immune deficiencies, not only bacterial infections but also the very serious opportunistic infections occur. Opportunistic lung infections are predominantly caused by cytomegalovirus (forming important combination with EBV in Table 3), two genera of fungi selected also in this paper, i.e. *Candida*, *Aspergillus*, and moreover by *Pneumocystis carinii* and fungi of genus *Mucor* (Ball, 1990). Similarly, organ transplantation can cause invasive mucosal infection such as aspegilloma, candidiasis and less frequently also similar infection with *Cryptococus neoformans* (Pacholczyk et al., 2011).

It is a question, what is the cause of superior occurrence of *Staphylococus aureus* (mentioned above) in blood cultures of PAH patients (49 of 192 patients) with central venous catheters (Oudiz et al., 2004). Though DQPA LXVNVT is responsible only for low positivity of IPAH-related EDO value in case of *S. aureus*, this structure is present in sequence of pathogenically important molecule of staphylokinase. This molecule abolishes bactericidal properties of α-defensins and activates plasmin, a broad-spectrum proteolytic enzyme facilitating bacterial penetration into the surrounding tissues and alleviates the

course of *S. aureus* sepsis (Bokarewa et al., 2006; Kwieciński et al., 2010). In addition, the sequence of other pathogenically important protein clumping factor B (responsible for nasal colonization with *S. aureus* (O'Brien et al., 2002; Walsh et al., 2009) contains other DQPA PXPPVXPXP. In case of the detection of the corresponding autoantibody specificities, the preceding facts would raise the question of possible inclusion of expression-related parameter to formula 2.

Web page references

1. Adzhubei AA, Sternberg MJ. Left-handed polyproline II helices commonly occur in globular proteins. J Mol Biol. 1993; 229:472-93. [PubMed: 8429558]
2. Adzhubei AA, Sternberg MJ. Conservation of polyproline II helices in homologous proteins: implications for structure prediction by model building. Protein Sci. 1994; 3:2395-410. [PubMed: 7756993]
3. Afonin PV, Fokin AV, Tsygannik IN, Mikhailova IY, Onoprienko LV, Mikhaleva II,Ivanov VT, Mareeva TY, Nesmeyanov VA, Li N, Pangborn WA, Duax WL, Pletnev VZ. Crystal structure of an anti-interleukin-2 monoclonal antibody Fab complexed with an antigenic nonapeptide. Protein Sci. 2001; 10:1514-21. [PubMed: 11468348]
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403-10. [PubMed: 2231712]
5. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389-402. [PubMed: 9254694]
6. Azevedo Sias S, Oliveira Caetano R, Dutra Comarella J, de Oliveira E, Santos Ferreira A, Quirico-Santos T. Successful treatment of lipoid pneumonia associated with bowel obstruction by Ascaris lumbricoides. J Trop Pediatr. 2009 [Epub ahead of print; PubMed: 20026557]
7. Ball F. X-ray diagnosis of immunologically induced lung diseases in children and adolescents. Radiologe 1990; 30:303-9. Review. [PubMed: 2205884].
8. Beardsley JR, Williamson JC, Johnson JW, Ohl CA, Karchmer TB, Bowton DL. Using local microbiologic data to develop institution-specific guidelines for the treatment of hospital-acquired pneumonia. Chest 2006; 130:787-93. [PubMed: 16963676]
9. Bills ND, Hinrichs SH, Morse JW. Direct detection of Epstein-Barr viral antigen in nasopharyngeal swabs from patients with infectious mononucleosis. Acad Emerg Med. 1996; 3:776-81. [PubMed: 8853673]
10. Bokarewa MI, Jin T, Tarkowski A. Staphylococcus aureus: Staphylokinase. Int J Biochem Cell Biol. 2006; 38:504-9. Review. [PubMed: 16111912]

11. Chiang YC, Lin YJ, Horng JC. Stereoelectronic effects on the transition barrier of polyproline conformational interconversion. Protein Sci. 2009; 18:1967-77. [PubMed: 19609932]
12. Chien JW, Johnson JL. Viral pneumonias. Infection in the immunocompromised host. Postgrad Med. 2000; 107:67-70, 73-4, 77-80. [PubMed: 10689409]
13. Cianciotto NP, Eisenstein BI, Mody CH, Toews GB, Engleberg NC. A *Legionella pneumophila* gene encoding a species-specific surface protein potentiates initiation of intracellular infection. Infect Immun. 1989; 57:1255-62. [PubMed: 2925251]
14. Conces DJ Jr. Endemic fungal pneumonia in immunocompromised patients. J Thorac Imaging. 1999; 14:1-8. [PubMed: 9894949]
15. Dagenais TR, Keller NP. Pathogenesis of Aspergillus fumigatus in Invasive Aspergillosis. Clin Microbiol Rev. 2009; 22:447-65. Review. [PubMed: 19597008]
16. De Keyser F, Hoch SO, Takei M, Dang H, De Keyser H, Rokeach LA, Talal N. Cross-reactivity of the B/B' subunit of the Sm ribonucleoprotein autoantigen with proline-rich polypeptides. Clin Immunol Immunopathol. 1992; 62:285-90. [PubMed: 1371727]
17. Dubaniewicz A. Mycobacterium tuberculosis heat shock proteins and autoimmunity in sarcoidosis. Autoimmun Rev. 2010; 9:419-24. [PubMed: 19931650]
18. Greene CM, Low TB, O'Neill SJ, McElvaney NG. Anti-proline-glycine-proline or antielastin autoantibodies are not evident in chronic inflammatory lung disease. Am J Respir Crit Care Med. 2010; 181:31-5. [PubMed: 19762563]
19. Hermans PW, Adrian PV, Albert C, Estevão S, Hoogenboezem T, Luijendijk IH, Kamphausen T, Hammerschmidt S. The streptococcal lipoprotein rotamase A (SlrA) is a functional peptidyl-prolyl isomerase involved in pneumococcal colonization. J Biol Chem. 2006; 281:968-76 [PubMed: 16260779]
20. Hirshberg B, Sklair-Levi M, Nir-Paz R, Ben-Sira L, Krivoruk V, Kramer MR. Factors predicting mortality of patients with lung abscess. Chest 1999; 115:746-50. [PubMed: 10084487]
21. Homolka S, Niemann S, Russell DG, Rohde KH. Functional genetic diversity among Mycobacterium tuberculosis complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. PLoS Pathog. 2010; 6:e1000988. [PubMed: 20628579]
22. Horiuchi K, Weskamp G, Lum L, Hammes HP, Cai H, Brodie TA, Ludwig T, Chiusaroli R, Baron R, Preissner KT, Manova K, Blobel CP. Potential role for ADAM15 in pathological neovascularization in mice. Mol Cell Biol. 2003; 23:5614-24. [PubMed: 12897135]
23. Huang L, Crothers K. HIV-associated opportunistic pneumonias. Respirology 2009; 14:474-85 [PubMed: 19645867]
24. Humbert M, Monti G, Brenot F, Sitbon O, Portier A, Grangeot-Keros L, Duroux P, Galanaud P, Simonneau G, Emilie D. Increased interleukin-1 and interleukin-6 serum concentrations in

severe primary pulmonary hypertension. Am J Respir Crit Care Med. 1995; 151:1628-31. [PubMed: 7735624]

25. Jakob RP, Zoldák G, Aumüller T, Schmid FX. Chaperone domains convert prolyl isomerases into generic catalysts of protein folding. Proc Natl Acad Sci U S A. 2009; 106:20282-7. [PubMed: 19920179]
26. Jura P. Fuzzy systems. In: Vavrin P, editor. Fundaments of fuzzy logics in control and modeling, Brno, Vutium, 2003, p. 62-74.
27. King AS, Castro JG, Dow GC. Nocardia farcinica lung abscess presenting in the context of advanced HIV infection: Spontaneous resolution in response to highly active antiretroviral therapy alone. Can J Infect Dis Med Microbiol. 2009; 20:e103-6. [PubMed: 20808449]
28. Königs C, Rowley MJ, Thompson P, Myers MA, Scealy M, Davies JM, Wu L, Ditrich U, Mackay CR, Mackay IR. Monoclonal antibody screening of a phage-displayed random peptide library reveals mimotopes of chemokine receptor CCR5: implications for the tertiary structure of the receptor and for an N-terminal binding site for HIV-1 gp120. Eur J Immunol. 2000; 30:1162-71. [PubMed: 10760806]
29. Kubrycht J, Borecký J, Soucek P, Ježek P, Ruzicka M, Sigler K. Recapitulation and improvement of our sequence approaches published in years 2002 and 2004. Prague, 2006. www.papersatellitesjk.com.
30. Kubrycht J, Sigler K. 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. 2008; 255:8-15. [PubMed: 18723029]
31. Kwieciński J, Josefsson E, Mitchell J, Higgins J, Magnusson M, Foster T, Jin T, Bokarewa M. Activation of plasminogen by staphylokinase reduces the severity of Staphylococcus aureus systemic infection. J Infect Dis. 2010; 202:1041-9. [PubMed: 20726765.
32. Lepš J. Binomial distribution. In: Pešek P, editor. Biostatistics. České Budějovice: University of Southern Bohemia, 1996, p. 145-8.
33. MacKay DN. Treatment of acute bronchitis in adults without underlying lung disease. J Gen Intern Med. 1996; 11:557-62. [PubMed: 8905509]
34. Marrie TJ, Fine MJ, Obrosky DS, Coley C, Singer DE, Kapoor WN. Community-acquired pneumonia due to Escherichia coli. Clin Microbiol Infect. 1998; 4:717-723. [PubMed: 11864280]
35. O'Brien LM, Walsh EJ, Massey RC, Peacock SJ, Foster TJ. Staphylococcus aureus clumping factor B (ClfB) promotes adherence to human type I cytokeratin 10: implications for nasal colonization. Cell Microbiol. 2002; 4:759-70. [PubMed: 12427098]
36. Mori T, Ebe T, Takahashi M, Isonuma H, Ikemoto H, Oguri T. Lung abscess: analysis of 66 cases from 1979 to 1991. Intern Med. 1993; 32:278-84. [PubMed: 8358116]

37. Morse JH, Barst RJ, Fotino M, Zhang Y, Flaster E, Fritzler MJ. Primary pulmonary hypertension: immunogenetic response to high-mobility group (HMG) proteins and histone. Clin Exp Immunol. 1996; 106:389-95. [PubMed: 8918589]
38. Niederman MS, Craven DE, Bonten MJ, Chastre J, Craigh VA, Fagon JY, Hall J, Jacoby GA, Kollef MH, Luna CM, Mandell LA, Torres A, Wunderink RG. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. Am J Respir Crit Care Med. 2005; 171:388-416. [PubMed: 15699079]
39. Oda Y, Okada Y, Katsuda S, Nakanishi I. Immunohistochemical study on the infection of herpes simplex virus, human cytomegalovirus, and Epstein-Barr virus in secondary diffuse interstitial pneumonia. Hum Pathol. 1994; 25:1057-62. [PubMed: 7927310]
40. O'Reilly LM, Daborn CJ. The epidemiology of Mycobacterium bovis infections in animals and man: a review. Tuber Lung Dis. 1995; 76 (Suppl 1): 1–46. [PubMed: 7579326]
41. Oudiz RJ, Widlitz A, Beckmann XJ, Camanga D, Alfie J, Brundage BH, Barst RJ. Micrococcus-associated central venous catheter infection in patients with pulmonary arterial hypertension. Chest 2004; 126:90-4. [PubMed: 15249447]
42. Pacholczyk M, Lagiewska B, Lisik W, Wasiak D, Chmura A. Invasive fungal infections following liver transplantation - risk factors, incidence and outcome. Ann Transplant. 2011; 16:14-6. [PubMed : 21959504]
43. Ren P, Rossettini A, Chaturvedi V, Hanes SD. The Ess1 prolyl isomerase is dispensable for growth but required for virulence in *Cryptococcus neoformans*. Microbiology 2005; 151:1593-605. [PubMed: 15870468]
44. Saubolle MA. Fungal pneumonias. Semin Respir Infect. 2000; 15:162-77. [PubMed: 10983933]
45. Schmid FX. Prolyl isomerase: enzymatic catalysis of slow protein-folding reactions. Annu Rev Biophys Biomol Struct. 1993; 22:123-42. [PubMed: 7688608]
46. Vázquez F, Hastings G, Ortega MA, Lane TF, Oikemus S, Lombardo M, Iruela-Arispe ML. METH-1, a human ortholog of ADAMTS-1, and METH-2 are members of a new family of proteins with angio-inhibitory activity. J Biol Chem. 1999; 274:23349-57. [PubMed: 10438512]
47. Waldman A, Gilhar A, Duek L, Berdicevsky I. Incidence of Candida in psoriasis--a study on the fungal flora of psoriatic patients. Mycoses 2001; 44:77-81. [PubMed: 11413927]
48. Walsh EJ, O'Brien LM, Liang X, Hook M, Foster TJ. Clumping factor B, a fibrinogen-binding MSCRAMM (microbial surface components recognizing adhesive matrix molecules) adhesin of Staphylococcus aureus, also binds to the tail region of type I cytokeratin 10. J Biol Chem. 2004; 279:50691-9. [PubMed: 15385531]
49. Wang XJ, Etzkorn FA. Peptidyl-prolyl isomerase inhibitors. Biopolymers 2006; 84:125-46. [PubMed: 16302169]

50. Yildiz O, Doganay M. Actinomycoses and Nocardia pulmonary infections. Curr Opin Pulm Med. 2006; 12:228-34. [PubMed: 16582679]
51. Zadik Y, Burnstein S, Derazne E, Sandler V, Ianculovici C, Halperin T. Colonization of Candida: prevalence among tongue-pierced and non-pierced immunocompetent adults. Oral Dis. 2010; 16:172-5. [PUMED: 19732353]
52. Zimmer R, Probabilistic Modeling, Orthodox and Bayesian Modeling HMMs. In: Algorithmische Bioinformatik II, Part II. 2004. www2.bio.ifi.lmu.de/lehre/WS2004/VLG_Algo_2/Material/041125_XI_Prob_HMM.pdf.

Table WT1. DQPA and some their properties (supplement to Table 2)

Sequence	DQPA name[1]	Occurrence[2]		DQPAn[3]	L_i[4]	E_{min}[4]	τ_i[4]
Fundamental DQPA subset							
PQxPVPRP	Q1	SSA27		DQPA7-*	8	1.480E-5	3.30340E-6
PVPRPP	Q2	CHD4		DQPA6-*	6	2.011E-5	5.98508E-6
PPRVLP	Q3	PTPR2		DQPA6-*	6	2.698E-5	8.03069E-6
PxYPPPP	Q4	Annexin A11		DQPA6	7	4.927E-5	1.25714E-5
YPPPP	Q5 = Q4a	Annexin A11	2/1nro	DQPA5	5	6.072E-5	2.16921E-5
PPPPLxPxP	Q6	CHD3		DQPA7	9	8.144E-5	1.61632E-5
PPPLxPxPP	Q7	CHD3		DQPA7	9	8.144E-5	1.61632E-5
PxPPVLPxP	Q8	HMG TOX2		DQPA7-*	9	1.093E-4	2.16875E-5
FPPGPxI	Q9	TG		DQPA6*	7	1.190E-4	3.03690E-5
PPPPLxP	Q10 = Q6g	leiomodin-1	2/1nro	DQPA6	7	1.190E-4	3.03690E-5
VPRPP	Q11 = Q2a	SOX13	2/1nro	DQPA5	5	1.467E-4	5.24022E-5
PPPPL	Q12 = Q6a2	TSYL2	3/1nro	DQPA5	5	1.467E-4	5.24022E-5
PPGPQ	Q13	TOX2_HUMAN		DQPA5	5	1.968E-4	7.03124E-5
PVPPG	Q14	Annexin A11		DQPA5	5	2.641E-4	9.43441E-5
PPxFPP	Q15	TG		DQPA5*	6	5.103E-4	1.51907E-4
RPxVEY	Q16	Ku70		DQPA5	6	5.103E-4	1.51907E-4
PPxLPxPP	Q17	CHD3		DQPA6	8	6.761E-4	1.50951E-4
GRPxxPPP	Q18	UBF1		DQPA6	8	9.072E-4	2.02543E-4
PPVxxVPP	Q19	ODP2		DQPA6*	8	9.072E-4	2.02543E-4
PxPPVE	Q20	CHD4		DQPA5	6	9.187E-4	2.73491E-4
PVXRPP	Q21	CHD4		DQPA5	6	9.187E-4	2.73491E-4
PPGFXP	Q22	Filamin-B		DQPA5-*	6	9.187E-4	2.73491E-4
PGxQxPVP	Q23	Annexin A11		DQPA6-*	8	1.217E-3	2.71769E-4
PVPxGF	Q24	SOX13		DQPA5*	6	1.233E-3	3.66966E-4
PPVXPV	Q25	PTPRN		DQPA5	6	1.233E-3	3.66966E-4
PSRXPP	Q26	TNR6A		DQPA5	6	1.233E-3	3.66966E-4
QLXQSI	Q27	Golgin 8A		DQPA5	6	1.654E-3	4.92389E-4
PPPXLG	Q28	AKA12		DQPA5	6	1.654E-3	4.92389E-4
PxPxxPVEY	Q29	SOX13		DQPA6-*	9	2.067E-3	4.10237E-4
LxVNVT	Q30	TG		DQPA5*	6	2.219E-3	6.60680E-4
PQXPXPXPP	Q31	ODP2		DQPA6	9	2.774E-3	5.50449E-4
PPxxQxPVP	Q32	Annexin A11		DQPA6-	9	3.722E-3	7.38584E-4
GxPxxPPPP	Q33	CHD3 + leiomodin-1		DQPA6	9	4.994E-3	9.91020E-4
DQPA6 derivatives							
QxPVPRP	Q1a = Q34	**cf. Q1 (SSA27)**		**DQPA6 only** -*	7	1.190E-4	3.03690E-5
PxxPVPRP	Q1b = Q35			**-**	8	6.761E-4	1.50951E-4
PQxxVPRP	Q1c = Q36			-*	8	6.761E-4	1.50951E-4
PQxPxPRP	Q1d = Q37			*	8	5.039E-4	1.12500E-4
PQxPVxRP	Q1e = Q38				8	6.761E-4	1.50951E-4
PQxPVPxP	Q1f = Q39				8	6.761E-4	1.50951E-4
PQXPVPR	Q1g = Q40			-*	7	1.190E-4	3.03690E-5
PPPLxPxP	Q6a=Q7g=Q41	cf. Q6 and Q7 (CHD3)			8	6.761E-4	1.50951E-4
PxPPLxPxP	Q6b = Q42	**cf. Q6 (CHD3)**			9	3.722E-3	7.38584E-4
PPxPLxPxP	Q6c = Q43				9	3.722E-3	7.38584E-4
PPPxLxPxP	Q6d = Q44			-	9	3.722E-3	7.38584E-4
PPPPxxPxP	Q6e = Q45				9	2.774E-3	5.50449E-4
PPPPLxxxP	Q6f = Q46	CHD3 + TSYL2			9	3.722E-3	7.38584E-4
PPPPLxP	Q6g = Q10	leiomodin-1 + CHD3			7	1.190E-4	3.03690E-5
PPLxPxPP	Q7a = Q47	**cf. Q7 (CHD3)**			8	6.761E-4	1.50951E-4
PxPLxPxPP	Q7b = Q48			-	9	3.722E-3	7.38584E-4
PPxLxPxPP	Q7c = Q49			*	9	3.722E-3	7.38584E-4
PPPxxPxPP	Q7d = Q50	CHD3+TSYL2			9	2.774E-3	5.50449E-4
PPPLxxxPP	Q7e = Q51	**cf. Q7 (CHD3)**		*	9	3.722E-3	7.38584E-4
PPPLxPxxP	Q7f = Q52			-	9	3.722E-3	7.38584E-4
PPPLxPxP	Q7g=Q6a=Q41	cf. Q6 and Q7 (CHD3)			8	6.761E-4	1.50951E-4
PPVLPxP	Q8a = Q53	**cf. Q8 (HMG TOX2)**		-*	7	1.597E-4	4.07486E-5
PxxPVLPxP	Q8b = Q54			-*	9	4.994E-3	9.91020E-4
PxPxVLPxP	Q8c = Q55				9	4.994E-3	9.91020E-4
PxPPxLPxP	Q8d = Q56				9	3.722E-3	7.38584E-4
PxPPVxPxP	Q8e = Q57			*	9	3.722E-3	7.38584E-4
PxPPVLxxP	Q8f = Q58				9	4.994E-3	9.91020E-4
PxPPVLP	Q8g = Q59			-	7	1.597E-4	4.07486E-5
Dense DQPA5 derivatives found by independent ScanProsite search and composing only fundamental DQPA							
PVPRP	Q1a2 = Q60	CHD4 + SSA27		**DQPA5 only**	5	1.467E-4	5.24022E-5
PPRVL	Q3a = Q61	**cf. the first section**			5	1.968E-4	7.03124E-5
PRVLP	Q3b = Q62	**of this table**			5	1.968E-4	7.03124E-5
PPVLP	Q8a2 = Q63				5	1.968E-4	7.03124E-5
FPPGP	Q9a = Q64			*	5	1.467E-4	5.24022E-5
DQPA8 of window extend W=10 not included to our set of nonapeptide epitope related structures							
PPPPLxPxPP	Q100	includes Q6 and Q7		DQPA8	10	9,570E-6	1,70932E-6

[1] **Code of DQPA names.** Order-related number of each DQPA is located after the Q character, forming order-related DQPA name. In addition to these order-related names, derivatives of fundamental DQPA are also associated with linkage-characterizing (**LC**) names. LC names are composed of i) order-related name of fundamental DQPA, which includes a sequence of given shorter derivative and ii) order character of such derivative (e.g. Q1b). In rare cases of two-times shortened derivatives, the numeral two terminates each LC name (e.g. Q8a2).

[2] number/1nro – though overall occurrence of DQPA identities in huAG sequences is higher than one, but only single occurrence is non-redundant when considering longer DQPA. For other abbreviations see Table 1.

[3] DQPAn – n means number of regular non-X DQPA; "-" denotes absence in Table WT2; * - absence in Table WT3.

[4] All minimum Expect values (E_{min}) were lower than 0.005 (i.e. $P < 0.005$; cf. WP7). τ_i - values of Expect terms necessary for EDO enumeration (cf. section 2.5). L_i - DQPA length.

Table WT2. Non-redundant occurrences of different DQPA in sequence sets of different animals forming S1.2 set[1] (supplement to Table 3)

DQPA [2]	QN	Q4	Q5	Q6	Q7	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q24	Q25	Q26	Q27	Q28	Q30	Q31	Q33	Q37	Q38	Q39	Q41	Q42	Q43	Q45	Q46	Q47	Q49	Q50	Q51	Q55	Q56	Q57	Q58	Q60	Q61	Q62	Q63	Q64
species [3]	N																																														
EBV (HHV-4)	1				1				3			2	1	1											1	2													1								
Nocardia farcinica	2																																						1								
Ascaris lumbricoides	3									1																														1							
CMV (HHV-5)	4				1		1		7		1						1			3	2												1							1							
Aspergillus clavatus	5		1					2	4			1								1	1					1				1		2	1			1			1				1			
Aspergillus fumigatus	6		1	1				2	8									1		1	2		1		1		1		1	1		2	1										1		
Mycoplasma pneumoniae	7																																							1						
St. aureus	8																					2		1																1						
P. aeruginosa	9					1			1															1									1				1				1			2	1		
M..tuberculosis	10	1	1						4	1	3				1		2	1	1							6						5				1				1			1	1		
Bordetella pertussis	11								1		1																												1		2	1	1		
Candida albicans	12		1					2									1		1			1		1		2						5		1			1			1					1		
Bl. dermatitidis	13																																								1						
Coccidioides immitis	14								3		2															2			1			1	3													
Klebsiella pneumoniae	15																1												1															1			
Herpes simplex Virus (HHV-1)	16			1				2	1					1												2																				
E. coli	17										3						3	1				3		2											1								1	1	1
Chl.pneumoniae	18																																1													
Histoplasma.capsulata	19										1															2							1													
HIV-1	20																																1				1									
hu adenovirus F	21															1																														
	QN	Q4	Q5	Q6	Q7	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q24	Q25	Q26	Q27	Q28	Q30	Q31	Q33	Q37	Q38	Q39	Q41	Q42	Q43	Q45	Q46	Q47	Q49	Q50	Q51	Q55	Q56	Q57	Q58	Q60	Q61	Q62	Q63	Q64

[1] WT2 in fact represents RSO set of hashes related to S1.2 subset of organisms in form of table. For the relationships between microorganism subsets and selection of non-redundant occurrences see sections 2.3 and 2.4, respectively. M. – *Mycobacterium*; P – *Pseudomonas*; S. – *Staphylococcus*. For other abbreviated names of organisms see Table 3.

[2] In accordance with Table 2 and WT1 some DQPA are not present in any of the compared sequence sets, i.e. the columns with such DQPA are not displayed here. QN – order related names of DQPA (cf. WT1).

[3] N – order related number of organism used in case of matrix representation in BASIC language derived program instead of hash key.

Table WT3. Non-redundant occurrences of different DQPA in sequence sets of different animals forming S2.2 set[1] (supplement to Table 3)

DQPA[2]	QN	Q4	Q5	Q6	Q7	Q10	Q11	Q12	Q13	Q14	Q16	Q17	Q18	Q20	Q21	Q25	Q26	Q27	Q28	Q31	Q32	Q33	Q35	Q38	Q39	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48	Q50	Q52	Q55	Q56	Q58	Q59	Q60	Q61	Q62	Q63
species[2]	N																																										
M. vanbaalenii	1					1		1			1			1																						1	1						
Yersinia enterocolitica	2													1																						1						2	
M. smegmatis	3													1																							1						
M. sp.	4							3						1	1																						1						
Cryptococus neoformans	5	1	2					16																				2		4	1		1		1			1			2	
FLV	6																																						1				1
SHFV	7																						1																2				
Helicobacter pylori	8																							1																			
HCV	9															1					1				1																		
HPV8	10			1				1																																			
Rubella virus	11				1							1					1																										
HHV-2	12							1												1						1				1													
Streptococcus gordonii	13																										1															
HPV47	14											1		1													1																
HPV5	15																	1				2					1																
HPV5b	16																	1				2					1																
Echinococcus granulosus	17																												1													
Trypanosoma brucei	18										1			1								2								1													
B-lymphotr. polyomavirus	19																													1													
M. bovis	20	1	1			1		4	1	3			1	2	1							4								4												1	
Rickettsia bellii	21																													1													
Ricket.tsia conorii	22							9											1											1												1	
Rickettsia montana	23							3																						1													
Rickettsia rickettsii	24							3											1											1													
HPV13	25																													1													
Brucella abortus	26																1													1												1	
Brucella ovis	27																													1												1	
Brucella Suis	28									1							1													1												1	
Salmonella dublin	29													1				1												1												1	
Rickettsia prowazekii	30																		1			2								1												1	
MMLV	31					1		1																																		
AMLV	32					1		1																																		
Neisseria meningitidis	33						2																									1									1	
Avian retrovirus	34																																	1								
Brucella melitensis	35																1													1												1	
HPV6a	36												1																				1	1								
HPV6b	37												1																					1								
HPV51	38								1										1																								
	QN	Q4	Q5	Q6	Q7	Q10	Q11	Q12	Q13	Q14	Q16	Q17	Q18	Q20	Q21	Q25	Q26	Q27	Q28	Q31	Q32	Q33	Q35	Q38	Q39	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48	Q50	Q52	Q55	Q56	Q58	Q59	Q60	Q61	Q62	Q63

[1] WT2 in fact represents ASO set of hashes related to S2.2 subset of organisms in form of table. For the relationships between microorganism subsets and selection of non-redundant occurrences see sections 2.3 and 2.4, respectively.

[2] AMLV, MMLV – avian and murine leukemia viruses; FLV – feline leukemia virus; HCV - hepatitis C virus; HHV-2 – human herpes virus 2; HPV – human papiloma virus; M - *Mycobacteruim*. For additional abbreviations see Table 3 and WT2. For additional comments see WT2.