BIOINFORMATICS DISCOVERY NOTE Vol. 22 no. 7 2006, pages 775–778
doi:10.1093/bioinformatics/btl006

Sequence analysis

Identification of a novel putative mitogen-activated kinase cascade on human chromosome 21 by computational approaches

Marialuisa Pellegrini-Calace[1] and Anna Tramontano[1,2,*]

[1]Department of Biochemical Sciences 'Rossi-Fanelli', University 'La Sapienza', Rome, Italy and
[2]Istituto Pasteur Fondazione 'Cenci-Bolognetti', Rome, Italy

Received on December 1, 2005; revised on December 29, 2005; accepted on January 13, 2006
Advance Access publication January 20, 2006
Associate Editor: Martin Bishop

ABSTRACT

Summary: Down syndrome (DS) is the most frequent form of mental retardation and is caused by chromosome 21 (HSA21) trisomy. Despite the number of known genes involved in DS and its high therapeutic interest, biological mechanisms leading to the DS phenotype are not fully clear. We present a functional hypothesis based on fold recognition and hidden Markov model techniques for four HSA21 genes located in the DS Candidate Region (DSCR). More specifically, we propose that they are members of a novel mitogen-activated protein kinase pathway with DYRK1A, SNF1LK and RIPK4 gene products being elements of the kinase cascade and the DSCR3 acting as structural scaffold for their interaction. This hypothesis finds support in various biochemical studies concerning the biological behavior and features of the involved HSA21 proteins. Our analysis calls for specifically designed experiments to validate our prediction and establish its relevance in terms of therapeutic approaches to the disease.

Contact: anna.tramontano@uniroma1.it

Suplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Down syndrome (DS) is the most frequent form of mental retardation and affects about 1 in 800 live births. DS is characterized by a well defined and distinctive phenotype, which includes various organ dysmorphies, stereotypic craniofacial anomalies and brain malformations (Gardiner *et al*., 2005; Epstein *et al*., 1991), and is caused by partial or full trisomy of chromosome 21 (HSA21). In particular, its peculiar cognitive and behavioural deficits are assumed to be because of the overexpression of some genes located on the HSA21 distal arm, specifically within the locus 21q22, in a region including about 20 genes and called DS Candidate Region (DSCR) (Delabar *et al*., 1993). HSA21 is among the shortest human chromosomes (~47 mbps) and, according to ENSEMBL, includes 271 genes, 243 of which are annotated as known genes, and 87 pseudogenes (Hubbard *et al*., 2005; Gardiner *et al*., 2003). Only two of the known genes are located on the HSA21 frontal arm, while 209 are found on the locus 21q22 (Pruitt *et al*., 2005). Despite the knowledge of supposedly all genes involved in DS, the high therapeutic interest of the disease and recent efforts aimed at helping the DS and HSA21 community in collecting and analysing available data (Nikolaienko *et al*., 2005), the biological mechanisms leading to the DS phenotype have not yet been fully clarified. In this work, we present a functional hypothesis for four HSA21 genes and in particular their identification as putative members of a novel mitogen-activated protein kinase (MAPK) pathway by fold recognition and hidden Markov model (HMM) approaches.

*To whom correspondence should be addressed.

2 RESULTS AND DISCUSSION

As a very first step, we focused on all HSA21 genes found in the OMIM database (Hamosh *et al*., 2005), which includes human genes related to genetic disorders (113 genes). Subsequently, genes with a well defined Gene Ontology (GO) (Ashburner *et al*., 2000) function annotation were excluded while genes characterized by an 'unknown' GO molecular function were chosen as candidates for our study (seven genes, see Supplementary Material, Table 1). A preliminary analysis of the corresponding gene products showed that the N-terminal domain of the DSCR3 protein (DSCR3p) belongs to the PFAM (Bateman *et al*., 2002) family PF03643, defined as the family of vacuolar protein sorting-associated protein 26 (Vps26) while nothing is known about the remaining seven proteins. Members of the PF03643 PFAM family are involved in protein trafficking and assemble into a complex with at least four other vacuolar protein sorting-associated proteins (Haft *et al*., 2000) where they act as a structural scaffold for the assembly of the complex.

DSCR3p was also used as query in various sequence similarity and fold recognition searches performed by BLAST (Altschul *et al*., 1990) and mGenTHREADER (McGuffin and Jones, 2003; Rychlewski and Fischer, 2005). Interestingly, mGenTHREADER, one of the best known fold recognition method, identified the crystal structure of visual arrestin (PDB ID: 1CF1) (Berman *et al*., 2000; Hirsch *et al*., 1999) as a hit for DSCR3p with its highest confidence level (raw score 0.959, *E*-value 3×10^{-5}).

Similar to Vps26, also members of the arrestin family associate with other proteins and act as scaffolds for their complexes. In particular, two arrestins, named β-arrestins, act as scaffolds in complexes linking the seven-membrane-spanning receptors (7MSR) to a number of effector pathways such as the MAPK pathway (Lefkowitz and Whalen, 2004). In the latter case they form complexes with the three final components of the pathway cascade (MAPKs, MAPK kinases and MAPK kinase kinases) allowing 7MSR to bind to the complex. Interestingly, although the function of arrestins and Vps26 proteins is very different, the PFAM

classification places the PF03643 family within the clan of the arrestin N-terminal domain, pointing at a structural relationship between the two classes of proteins.

Moreover, mGenTHREADER predicts that the structure similarity between DSCR3p and the arrestin fold is not limited to the N-terminal domain but involves the whole protein: the alignment contains 275 positions (the lengths of DSCR3p and of visual arrestin are 373 and 297, respectively).

The similarity of the scaffolding role of β-arrestins and Vps26 in their respective complexes and the structure similarity between DSCR3p and the arrestin fold suggests that DSCR3p could act as a scaffold element in a MAPK-like pathway. The observation that MAPK pathways are responsible for pivotal cellular processes and that the DS phenotype is a gene-dosage dependent event prompted us to investigate whether genes encoding for members of a MAPK cascade that could use DSCR3p as scaffold are located on human chromosome 21. MAPKs are the main cellular elements that control the communications of extracellular signals into the nucleus and the subsequent regulation of gene expression and are therefore involved in a number of fundamental cellular processes, such as apoptosis, cell proliferation, cell motility and cell differentiation. MAPKs are activated by dual-phosphorylation on threonine and tyrosine residues on a TXY motif within the activation loop by a signal module involving a MAPK-kinase (MKK) and a MAPK-kinase-kinase (MKKK) (Whitmarsh and Davis, 1998).

HSA21 can express six proteins belonging to the broad class of kinases: the dual-specificity tyrosine-phosphorylated and regulated kinase 1A (DYRK1A), the hormonally upregulated Neu-associated kinase (HUNK), a pyridoxal kinase (PDXK), a phosphofructokinase (PFKL), the receptor-interacting serine–threonine kinase 4 (gene RIPK4, gene product RIP4) and a sucrose non-fermenting 1 (SNF1) like protein kinase (SNF1LK) (Hattori *et al*., 2000). The exact cellular function of four of them, DYRK1A, HUNK, RIPK4 and SNF1LK, is still poorly defined. DYRK1A is particularly interesting: it maps to the DSCR (locus 21q22.2), and it has been shown to be distantly related to MAP kinases (Miyata and Nishida, 1999; Galceran *et al*., 2003) and to enhance the MAPK cascade. It is transiently translocated to the nucleus, but soon after is transported to the growing dendritic tree where it is supposed to phosphorilate its substrates (Hammerle *et al*., 2003). In fact, in PC12 cells, it can form a multimeric complex with Ras, B-Raf and MEK1 (Kelly and Rahmani, 2005), components of the signal module activating either one of the extracellular-signal-regulated kinases 1 and 2 (ERK1/ERK2), one of the four main MAPKs hitherto identified.

In order to assess whether any of the HSA21 kinases, or any of the other proteins encoded by chromosome 21, could be the kinases forming a complex with DSCR3p, a HMM was derived for each of the three last components of a canonical MAPK cascade (MAPK, MKK and MKKK) and used to scan the set of all HSA21 encoded proteins.

2.1 Derivation and validation of HMMs

HMMs were derived using a homology-based sequence training set composed by three sub-sets, one for each of the three considered MAPK cascade elements. The human sequences of the ERK1/2 module ERK2, MEK1 and RAF1 (UniProt code MK01_HUMAN, MP2K1_HUMAN and RAF1_HUMAN, respectively) were scanned against the UniRef90 database by a BLAST search and the retrieved sequences were included into the corresponding subset.

The HMM's discriminating ability was assessed and optimized using a validation set of MAPK pathway members annotated as such in the UniProt database that were not identified by BLAST at an expectation value (*E*-value) threshold of 10^{-6} and <1000 amino acids in length (74 sequences: 34 MAPKs, 14 MKKs and 26 MKKKs). This threshold was necessary because the length of MKKKs is quite variable, significantly more than that of MAPKs and MKKs. In fact, MKKKs have a minimum length of 468 amino acids, a maximum length of 1608 amino acids and a standard deviation of 361.5 amino acids, while the corresponding values for MAPKs are 275, 807 and 111 and those for MKKs are 63, 449 and 88.5.

It should be mentioned here that there is a certain degree of philogenetic promiscuity between the MKKs and MKKKs classes of kinases. They both are serine–threonine protein kinases and are classified in the same kinase group (STE) by Manning and coworkers (Manning *et al*., 2002). Accordingly, the guide tree generated by CLUSTALW for the 74 MAPK-involved sequences in the test set (see Supplementary Material, Figure 1) shows that MKKKs do not fall into a unique branch as MAPKs and MKKs do, but generate three different and distant clusters, one of them closer to the MKKs branch than to those of the remaining MKKKs. These data, together with the discussed variability in the length of MKKKs, led us to limit our analysis to the kinase domain in this case, in order to avoid the model to be biased by constraints other than their phosphorylation target specificity (e.g. differences in the protein interaction pattern). The MKKK–HMM therefore only includes the sequence of the kinase domain (amino acids 349–609, numbering of RAF1_HUMAN).

The validation was carried out as follows: all BLAST hits with an *E*-value lower than an initial threshold of 10^{-6} were collected, aligned by ClustalW1.83 (Thompson *et al*., 1994) and subsequently used to derive the corresponding HMM using HMMER2.3.2 (Eddy, 1998). The discriminative power of each resulting HMM was estimated on the validation set in terms of its ability to discriminate between sequences belonging to the targeted group of kinases and sequences belonging to the other two groups. If the HMM discriminative power resulted poor, the BLAST *E*-value threshold was decreased and the procedure repeated. This validation cycle was re-iterated until the three HMMs discriminating the highest number of kinase sequences in the targeted group were found.

MAPK-HMM was derived using an *E*-value threshold of 10^{-46}. Its bit score was always >776.4 for MAPK sequences and never >−33.4 for kinases of the the MKKs and MKKKs groups. *E*-values for MAPK sequences were all $<10^{-93}$. *E*-values for the other groups were also significant (between 10^{-13} and 10^{-6}), reflecting the similarity of their molecular function.

MKK–HMM and MKKK–HMM were derived with a threshold of 10^{-65} and 10^{-50}, respectively. MKK–HMM could discriminate MKK sequences (bit-scores above 283.2 and *E*-values lower than 10^{-83} in all cases) and only two of the MKKKs had a positive, or nearly positive, bit-score when compared with the model (M3K2_ARATH: bit score 7.7 and M3K1_ARATH: bit score of −1.3). MKKK–HMM gave a positive bit score for 16 of the 26 MKKKs sequences in the set with all the corresponding *E*-values $<10^{-10}$.

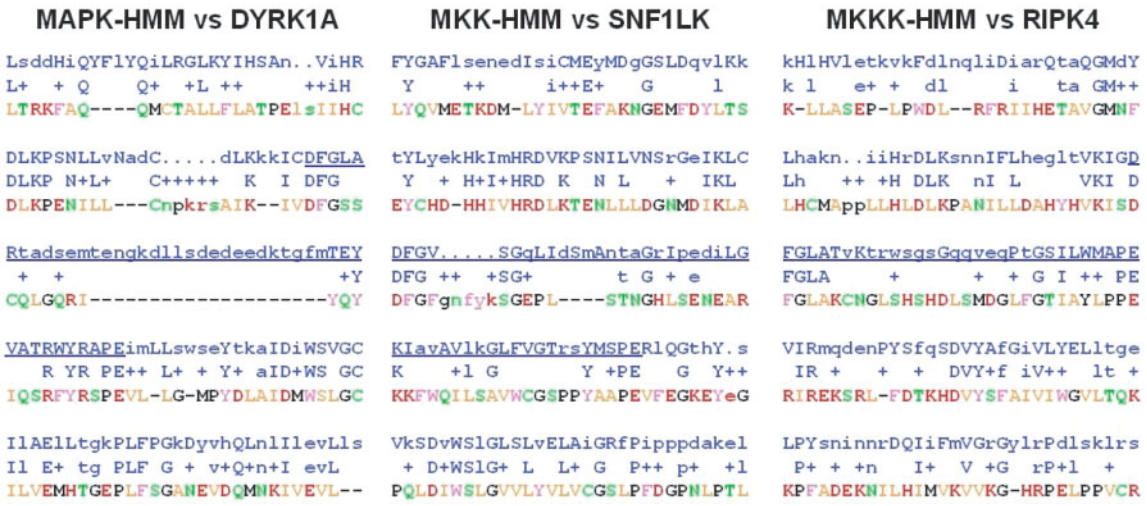

Fig. 1. Pairwise alignments of part of the kinase domain of DYRK1A, SNF1LK and RIPK4 to MAPK–HMM, MKK–HMM and MKKK–HMM (bit scores = 70.5, −97.5, 36.7; *E*-values = 10^{-18}, 10^{-8}, 2.4×10^{-11}, respectively). The HMM consensus is in blue, the colour code of the aligned sequence is red for charged amino acids, green for polar amino acids, orange for non-polar amino acids and magenta for aromatic amino acids. The kinase motif is underlined in the HMM consensus.

2.2 MAPK cascade-HMMs versus HSA21

Once obtained, our final HMMs were used to search against the 271 HSA21 protein sequences representing the products of all known genes of the chromosome, in order to identify possible members of a putative MAPK-like signal pathway.

MAPK–HMM produced a positive bit score only for DYRK1A and MKKK-HMM only for RIPK4. The best hit for MKK–HMM was SNF1LK, although in this case the bit score was negative (−97.5 and *E*-value of 10^{-8}).

Pairwise alignments of a fragment of the kinasic domain, nearby the activation segment, of DYRK1A, SNF1LK and RIPK4 to MAPK-HMM, MKK-HMM and MKKK-HMM respectively are shown in Figure 1. As shown in the figure, all three proteins show an overall good sequence similarity with the corresponding HMM-consensus near the activation segment. The number of aligned amino acids is 371 for DYRK1A and RIPK4 and 296 for SNF1LK, to be compared with a total protein sequence length of 540, 784 and 783, respectively.

The MAPK–HMM/DYRK1A alignment, which covers more than two-third of the protein sequence, matches the dual-tyrosine motif YXY with the TXY of classical MAPKs and shows a long deletion corresponding to a region which is very variable in length among the MAPK sequences. SNF1LK and RIPK4 matched MKK/HMM and MKKK/HMM only with their kinase domain, i.e. with about a half of their sequences.

Interestingly, the non-kinase domain of the RIPK4 gene product (known as RIP4, DIK or PKK) is formed by 11 C-terminal ankyrin repeats.

Experimental studies have shown that RIP4 potently activates the JNK pathway, one of the four known MAPK pathways, in a kinase-domain dependent way, and homodimerises and is probably negatively regulated via both the kinase and the ankyrin domains. The latter are among the most common protein–protein interaction motifs (Meylan *et al*., 2002; Meylan and Tschopp, 2005; Mosavi *et al*., 2004) and have been shown to interact *in vitro* with

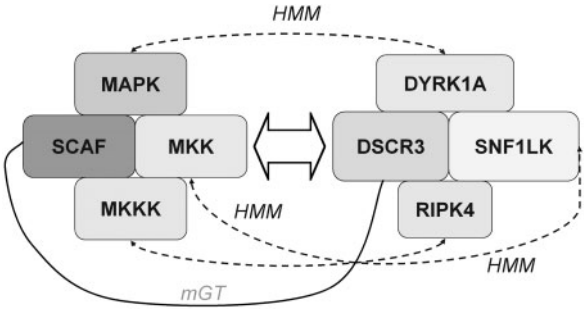

Fig. 2. Schematic representation of a generic mitogen activated cascade and of the proposed HSA21 MAPK cascade. MAPK, MKK, MKKK and SCAF indicate general element of the pathway. Dotted blue lines represent predictions obtained by HMMs; the solid line indicates the mGenTHREADER (mGT) fold recognition prediction.

serine–threonine kinases (Lin *et al*., 1999) and to be involved to some extent in the JNK pathway. In particular, AKRL1 and AKRL2, two non-kinase proteins containing several ankyrin repeats and transmembrane domains, have been shown to activate JNK and its activators MKK4 and MKK7 and to interfere at the MKKK level with an as yet unclear mechanism (Harada *et al*., 2003).

Results obtained by the HMM predictions and fold recognition, along with previously published experimental data (Kelly and Rahmani, 2005; Meylan and Tschopp, 2005), suggest that a MAPK-like pathway composed by DYRK1A as MAPK, SNF1LK as MKK, RIP4 as MKKK and DSCR3p as scaffold protein might be encoded by HSA21 (Fig. 2). The hypothesis is further supported by a recently published experimental study performed on postmortem brain tissue from Alzheimer's disease and DS patients, which showed that MAP kinase activity is increased in both pathologies (Swatton *et al*., 2004). Moreover, HSA21 also encodes for the transcription factor ETS2 (Baldus *et al*., 2004), known to be phosphorylated upon activation of the MAPK pathway and to be

involved in the development of the acute myeloid leukaemia, which has an incidence of 10- to 20-folds higher in DS children (Gurbuxani *et al*., 2004). Classical genetic studies involving aneuploidy and ploidy comparisons showed changes of single chromosomal dosage to alter the phenotype more than changes in ploidy. Recent evidences suggest that the behaviour of molecular complexes functioning in regulatory processes contributes importantly to this balance (Birchler *et al*., 2005). It seems therefore reasonable to hypothesize that, in DS patients, the overexpression of our putative MAPK-like pathway caused by the trisomy could lead to the over-activation of its effectors and have a role in the disruption of the fundamental cell processes in which known MAPK pathways are involved [DYRK1A has already been shown to be involved in DS mental retardation (Branchi *et al*., 2004)]. We trust that the data presented here will encourage the design of experiments aimed at testing the ability of DYRK1A, SNF1LK, RIP4 and DSCR3p to interact and act *in vivo* as a mitogen activated kinase cascade, activating the 7MSR effector pathways.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Veronica Morea for useful comments. This work was supported by the BioSapiens Network of Excellence, European Commission FP6 Programme, contract number LHSG-CT-203-503265 and by the Istituto Pasteur, Fondazione Cenci Bolognetti. M.P.C. gratefully acknowledges the support from 'FIRB 2001: Bioinformatica per la Genomica e la Proteomica'.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Baldus,C.D. *et al*. (2004) Acute myeloid leukemia with complex karyotypes and abnormal chromosome 21: Amplification discloses overexpression of APP, ETS2, and ERG genes. *Proc. Natl Acad. Sci. USA*, **101**, 3915–3920.

Bateman,A. *et al*. (2002) The Pfam protein families database. *Nucleic Acids Res*., **30**, 276–280.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.

Birchler,J.A. *et al*. (2005) Dosage balance in gene regulation: biological implications. *Trends Genet*., **21**, 219–226.

Branchi,I. *et al*. (2004) Transgenic mouse *in vivo* library of human Down syndrome critical region 1: association between DYRK1A overexpression, brain development abnormalities, and cell cycle protein alteration. *J. Neuropathol. Exp. Neurol*., **63**, 429–440.

Delabar,J.M. *et al*. (1993) Molecular mapping of twenty-four features of Down syndrome on chromosome 21. *Eur. J. Hum. Genet*., **1**, 114–124.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Epstein,C.J. *et al*. (1991) Protocols to establish genotype–phenotype correlations in Down syndrome. *Am. J. Hum. Genet*., **49**, 207–235.

Galceran,J. *et al*. (2003) The MNB/DYRK1A protein kinase: genetic and biochemical properties. *J Neural Transm Suppl*., **67**, 139–148.

Gardiner,K. *et al*. (2003) Mouse models of Down syndrome: how useful can they be? Comparison of the gene content of human chromosome 21 with orthologous mouse genomic regions. *Gene*, **318**, 137–147.

Gardiner,K. *et al*. (2005) Report on the 'Expert Workshop on the Biology of Chromosome 21: towards gene-phenotype correlations in Down syndrome', held June 11–14, 2004, Washington D.C. *Cytogenet. Genome Res*., **108**, 269–277.

Gurbuxani,S. *et al*. (2004) Recent insights into the mechanisms of myeloid leukemogenesis in Down syndrome. *Blood*, **103**, 399–406.

Haft,C.R. *et al*. (2000) Human orthologs of yeast vacuolar protein sorting proteins Vps26, 29, and 35: assembly into multimeric complexes. *Mol. Biol. Cell*, **11**, 4105–4116.

Hammerle,B. *et al*. (2003) Expression patterns and subcellular localization of the Down syndrome candidate protein MNB/DYRK1A suggest a role in late neuronal differentiation. *Eur. J. Neurosci*., **17**, 2277–2286.

Hamosh,A. *et al*. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.

Harada,T. *et al*. (2003) AKRL1 and AKRL2 activate the JNK pathway. *Genes Cells*, **8**, 493–500.

Hattori,M. *et al*. (2000) The DNA sequence of human chromosome 21 [Erratum (2000) *Nature*, **407**, 110.]. *Nature*, **18**, 311–319.

Hirsch,J.A. *et al*. (1999) The 2.8 Å crystal structure of visual arrestin: a model for arrestin's regulation. *Cell*, **97**, 257–269.

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.

Kelly,P.A. and Rahmani,Z. (2005) DYRK1A enhances the mitogen-activated protein kinase cascade in PC12 cells by forming a complex with Ras, B-Raf, and MEK1. *Mol. Biol. Cell*, **16**, 3562–3573.

Lefkowitz,R.J. and Whalen,E.J. (2004) beta-arrestins: traffic cops of cell signalling. *Curr. Opin. Cell Biol*., **16**, 162–168.

Lin,J.H. *et al*. (1999) The ankyrin repeat-containing adaptor protein Tvl-1 is a novel substrate and regulator of Raf-1. *J. Biol. Chem*., **274**, 14706–14715.

Manning,G. *et al*. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.

McGuffin,L.J. and Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, **19**, 874–881.

Meylan,E. *et al*. (2002) RIP4 (DIK/PIKK), a novel member of the RIP kinase family, activates NF-kB and is processed during apoptosis. *EMBO reports*, **3**, 1202–1208.

Meylan,E. and Tschopp,J. (2005) The RIP kinases: crucial integrators of cellular stress. *Trends Biochem. Sci*., **30**, 151–159.

Miyata,Y. and Nishida,E. (1999) Distantly related cousins of MAP kinase: biochemical properties and possible physiological functions. *Biochem. Biophys. Res. Commun*., **266**, 291–295.

Mosavi,L.K. *et al*. (2004) The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci*., **13**, 1435–1448.

Nikolaienko,O. *et al*. (2005) Human chromosome 21/Down syndrome gene function and pathway database. *Gene*, **364**, 90–98.

Pruitt,K.D. *et al*. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.

Rychlewski,L. and Fischer,D. (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci*., **14**, 240–245.

Swatton,J.E. *et al*. (2004) Increased MAP kinase activity in Alzheimer's and Down syndrome but not in schizophrenia human brain. *Eur. J. Neurosci*., **19**, 2711–2719.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Whitmarsh,A.J. and Davis,R.J. (1998) Structural organization of MAP-kinase signalling modules by scaffold proteins in yeast and mammals. *Trends Biochem. Sci*., **23**, 481–485.