Supplementary Methods for KM Gilljam et al, *Identification of a new and widespread PCNA binding motif* (2009)

Methods

Sequence analysis

The procedure used for sequence analysis of potential APIM motifs was implemented in the `confind` tool (Drabløs, unpublished), integrating the individual steps described below. Full output is available in the accompanying output files for APIM (`confind_APIM_out.html`, described here) and PIP (`confind_PIP_out.html`). For initial sequence analysis the Swiss-Prot and TrEMBL databases [1] were used to find proteins with similar sub-sequences as the sequence region of interest. The databases were queried with motifs in PROSITE format [2]. Clustal W [3] was used to align the sequences of interest. The conserved motifs listed in the output file were identified by comparison of gene orthologs. Data files for Inparanoid [4, 5] version 5.1 were downloaded from the Inparanoid web server <http://inparanoid.sbc.su.se/> for a representative subset of organisms (see output file for details). The human sequences were used as reference, and the Inparanoid processed fasta file was searched with a regular expression for the APIM motif, using a local tool. A slightly expanded motif definition was used, where Ala was allowed at either position 3 or 4 of the motif, in addition to Ile, Val and Leu, but not at both positions simultaneously. From a total of 22218 protein sequences there were 636 sequences with at least one hit against the APIM motif. These entries were matched against experimental and predicted subcellular localization in the eSLDB database [6], downloaded from the web server <http://gpcr.biocomp.unibo.it/esldb/>, and 349 entries with no indication of targeting to the nucleus were removed. For the remaining 287 entries the corresponding Inparanoid orthologs were identified, the corresponding sequences were extracted from the fasta files, and the resulting sequence libraries were aligned with Clustal W [3]. The 24 sequence entries without orthologs in Inparanoid were removed from the analysis. Two different procedures were used in

parallel for identification of conserved sites. In the first procedure (Consensus) the consensus sequence was estimated from the multiple alignments for each hit position in the human sequence. When estimating the consensus equivalent symbols in the conservative APIM motif (without Ala) where treated as equivalent symbols for estimation of the consensus, so that e.g. Ile, Val and Leu were treated as a single residue type. The regular expression was tested again against the consensus before the hit position was accepted. In the alternative procedure (Individual) the regular expression was tested against each orthologous subsequence corresponding to a hit position in the human sequence, and only positions where at least 50% of the orthologs matched the expression were accepted. In this estimate subsequences consisting only of gaps were excluded, assuming that this could represent e.g. alternative splice variants. These two procedures gave almost identical results, and the combined output is shown in the output file. In total 37 entries were removed by this procedure, the remaining 226 entries were listed and analyzed. The protein descriptions used in the output were taken from the Inparanoid unprocessed human fasta file and Ensembl [7] release 45. The output file is in html format and can be opened by a standard web browser.

References

1. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research *31*, 365-370.
2. de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res *34*, W362-365.
3. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (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.
4. Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol *314*, 1041-1052.
5. O'Brien, K.P., Remm, M., and Sonnhammer, E.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res *33*, D476-480.

6. Pierleoni, A., Martelli, P.L., Fariselli, P., and Casadio, R. (2007). eSLDB: eukaryotic subcellular localization database. Nucleic Acids Res *35*, D208-212.
7. Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. (2007). Ensembl 2007. Nucleic Acids Res *35*, D610-617.