

# Inference of Structure, Function and Motion of an Ancestral Protein

by Lee Pedersen, Jotun Hein and Mark Sansom.

An interested student should apply to this scheme: <http://oxcam.gpp.nih.gov/> with deadline 15.12.08. Please email any of the three researchers above for more information.

## Background and Motivation

Comparative Biology is a major contributor to biological understanding and can be applied to any biological objects that are homologous. The large application area presently is sequences, but other rising areas include structures, networks, organs and more. The strength of evolutionary comparison is the ability to detect features of functional importance and thus select properties that demands a functional explanation from a sea of noise. The only requirement for evolutionary comparison is that the objects to be compared are homologous. In this sense movements are perfectly homologous as the movements associated an ancestral molecule would be inherited with modification, when the molecule evolved over evolutionary time. The recent rise in the capability of molecular dynamics, the number of known extant structures and the ability to predict structures, allows increasingly ambitious projects to be undertaken. This project combines all these components to a specific data set, attempting to solve a specific problem: A protein (Protein Z (PZ)) is found in humans in substantial concentration that appears to have cofactor function. It is homologous to several other proteins, all serine proteases. It appears to have either lost enzymatic activity or to have never properly evolved to have such function. Inferring the ancestral protein and exploring its properties computationally could solve this question. The project would also lead to the development of generally applicable techniques.

## Proposed Research

The project would have following components: Data collection, sequence analysis, ancestral structure prediction, MD simulations and integrated analysis.

- Continued collection of structures, sequences and MD data on the set of proteins homologous to MZ.
- Sequence analysis can be done optimally with software such as StatAlign (Novak et al., 2008) that is based on explicit probabilistic models of sequence evolution and can give confidence intervals on the reliability of the reconstructed ancestral sequences. Additionally, rates and modes of selection can be inferred for individual residues.
- Ancestral Structure Prediction will in this case be done by homology modeling as we want to predict a structure for a reconstructed sequence when structures of closely related sequences are available.
- Molecular Dynamics Simulations will be done on the reconstructed ancestral protein to infer its dynamic properties.
- Integrated Analysis will combine sequence, structure and motion analysis to give a functional interpretation of the Z protein.

Protein Z (PZ) is a vitamin K-dependent (VKD) protein relatively abundant in humans, with wide plasma concentration range (Miletich and Broze 1987). PZ circulates in plasma as a complex with protein Z-dependent protease inhibitor (ZPI) (Tabatai et al. 2001). The most important known physiological function of PZ is its ability to enhance the inhibition of factor Xa by ZPI 1000-fold in the presence of phospholipid membrane and Ca<sup>2+</sup> ions (Al-Shanqeeti et al. 2005; Han et al. 1998). PZ is homologous to coagulation factors FVII, FIX, FX, and Protein C (PC), and the domain assembly of PZ is identical to these proteins. It is composed of four domains: a N-terminal Gla domain, two epidermal growth factor (EGF)-like domains (EGF1 and EGF2 domains) and a SP-like domain (Sejima et al. 1990; Ichinose et al. 1990). However, PZ lacks the critical histidine and serine residues of the catalytic triad, and is therefore not a zymogen of a serine protease. Based on its homology to the other coagulation factors, a solvent-equilibrated model of human PZ was proposed using comparative modeling and molecular dynamics (MD) simulation (Lee et al. 2007). One of the striking features in this model was that even though PZ lacked two of the critical catalytic triad residues, the relative spatial arrangement of the putative active site region and the distances between these residues were similar to the other serine proteases after long time simulation. The two  $\beta$ -barrel domains in PZ were juxtaposed similar to other chymotrypsin-like enzymes even without an active

catalytic triad bridging them. In addition, the residues Met309 and Val331, which correspond to a disulfide bond (Cys191-Cys220, chymotrypsinogen numbering) in other VKD SP-domains, were close to disulfide bond forming distance without distance or position constraints imposed. These observations raised questions about the evolution of PZ and led us consider if PZ could have once been functionally active as a serine protease.

Studies on serine protease evolution have shown that apart from protease domain sequences, the development and maintenance of the active site structure plays an important role in the evolution of proteolytic enzymes (Krem and Di Cera 2001, 2002). The preservation of the active site three-dimensional (3D) structure in PZ compared to other serine proteases indicates the operation of a possible divergent evolution mechanism within this structural family and raises the question of whether an active PZ was disfavored in evolution through sequence mutations in the protease domain. A possible reason for this evolutionary choice may have been due to negative evolutionary pressure brought about by undesirable function. Another possibility is that PZ might have been an ancestor for FVII, FIX, FX and PC. This possibility, however, seems unlikely based on genomic evidence from a recent study that characterized blood coagulation proteins (using new genetic databases) in a jawless vertebrate system (lamprey) (Doolittle et al. 2008). It was found that homologs for FVII, FIX, FX and PC exist in jawless vertebrates whereas no matches are found for PZ. The earliest homolog of PZ is found only in jawed vertebrates like fugu and zebrafish which indicates that PZ evolved later than the other coagulation serine proteases (Davidson et al. 2003).

In order to understand the evolutionary significance, the three dimensional structure of the putative activated form of PZ (PZa) was recently modeled (Chandrasekaran, 2008). Apart from the evolutionary aspect, this *in silico* model building is also interesting as a protein design challenge to generate an activated form of a serine protease starting from an inactive form. A recent experimental work (Ortlund et al. 2007) has shown that it is possible to resurrect an ancestral protein by employing evolutionary relationship to sequence comparisons, dramatically leading to the X-ray crystal structure. Given the existing 3D model for protein Z, an *in silico* PZa model was constructed by designing a primary sequence of PZa that might reasonably resurrect serine protease catalytic activity and then simulating a three-dimensional solvent-equilibrated structure of the activated form. Further a molecular docking study was employed to partially validate the predicted model by evaluating its ability to bind *in silico* to known serine protease inhibitors for FVIIa and FXa.

*Given this body of experimental evidence and given the theoretical modeling that tends to substantiate the similarities of Protein Z to other active serine proteases, how did Protein Z arrive in its current state.? Is it a precursor zymogen for an active PZa? Or is it a left-over of some evolutionarily-driven process.*

### **Timeline, Supervision and Collaboration**

The student will be jointly supervised by Lee Pedersen (NIEHS & UNC, North Carolina, USA), Jotun Hein (Oxford, UK) and Mark Sansom (Oxford, UK). The scholarship starts October 2009 and is of 4 years duration. A successful application will spend approximately equal amounts of time in US and UK. The project has great potential, but is also demanding as it covers Molecular Dynamics, Structure Prediction, Statistical Models of Evolution and also Software and Algorithm Design. Applicants must have strong qualifications in the relevant areas.

### **References**

- Al-Shanqeeti, et al. 2005. Protein Z and protein Z-dependent protease inhibitor *Thromb. Haemost.* **93**: 411-413
- Chandrasekaran, V. et al. Prot. Sci. (in press) 2008
- Davidson, C.J., Tuddenham, E.G., and McVey, J.H. 2003. 450 million years of hemostasis *J.Thromb. Haemost.* **1**: 1487-1494
- Doolittle, R.F., Jiang, Y. and Nand, J. 2008. Genomic evidence for a simpler clotting scheme in jawless vertebrates *J. Mol. Evol.* **66**: 185-196
- Han, X., Fiehler, R., and Broze, G.J. Jr. 1998. Isolation of a protein Z-dependent plasma protease inhibitor *Proc. Natl. Acad. Sci.* **95**: 9250-9255
- Ichinose, A. et al. E.W. 1990. Amino acid sequence of human protein Z, a vitamin K-dependent plasma glycoprotein *Biochem. Biophys. Res. Commun.* **172**:1139-1144
- Lee, C.J. et al. 2008 *J. Throm. Haemost.* **5**: 1558-1561
- Krem, M., and Di Cera, E. 2002. Evolution of enzyme cascades from embryonic development to blood coagulation *TRENDS in Biochemical Sciences* **27**: 67-74
- Miletich, J.P., and Broze, G.J. Jr. 1987. Human plasma protein Z antigen: range in normal subjects and effect of wafin therapy *Blood* **69**: 1580-1586
- Novak, A., Miklos, I., Lyngsø, R. & Hein, J. (2008) StatAlign: An extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*
- Ortlund, E.A. et al. 2007. Crystal structure of ancient protein: evolution by conformational epistasis *Science* **317**: 1544-1548
- Sejima, et al. 1990. Primary structure of vitamin K-dependent human protein Z *Biochem. Biophys. Res. Commun.* **171**: 661-668
- Tabatabai, A., Fiehler, R., and Broze, G.J. Jr. 2001. Protein Z Circulates in Plasma in a Complex with Protein Z-Dependent Protease Inhibitor *Thromb. Haemost.* **85**: 655-660