Abstract

Dipeptidyl peptidase 4 (DPP4) and its homologous family members DPP8 and DPP9 are a class of post-prolyl bond-cleaving serine exopeptidases. DPP4 is best known for its ability to rapidly degrade incretin peptides, glucagon-like peptide-1 and glucose-dependent insulinotropic polypeptides. DPP4 inhibitors are a class of popular insulinotropic agents without significant glycaemic potential, which are widely used in clinics as a second-line defence in managing type-2 diabetes mellitus. DPP4 catalytic activity has been associated with modulating immune response in the pathological context of obesity, type-2 diabetes, liver complications and many cancer types. Dysfunctional regulation of DPP4 may prolong inflammation, disrupt subclinical immune activities, and lead to metabolic imbalance. To further complicate the DPP4 regulatory landscape, there is a wealth of seemingly contradictory evidence to support opposing DPP4 functions in disease pathophysiology. This thesis aimed to develop a comprehensive understanding of the regulatory networks for DPP4 by establishing a complete perspective on DPP4 substrates.

This study identified a total of 147 novel DPP4 substrates through a combination of bioinformatics data mining (116 new substrates) and a custom-designed and trained hybrid neural network model (an additional 31 novel discoveries). The novel DPP4 substrates discovered are mainly found to come from six heavily networked pathophysiological functional groups, including immunity and digestion regulations, and more interestingly, DPP4 has also been found to have an in-depth connection in neural signalling conduction. This study further evaluated the potential for DPP4 to be a biomarker in pathogenic pain that may be a result of prolonged inflammation in the peripheral nervous system. Furthermore, DPP4's enzymatic property may also underpin DPP4’s regulatory effects in psychological conditions like depression and clinical substance addictions.

This study also attempted to develop a high-throughput low setup matrix-free top-down tandem mass spectrometry workflow to effectively verify small N-terminal dipeptide cleavage *in vivo.* An N-terminal dipeptide was successfully identified for cholecystokinin and supported *in silico* discovery of cholecystokinin as a novel physiological substrate of DPP4. Although more efforts are still required to better control sample degradation, this high-throughput workflow has the potential to provide a rapid pathway for small proteolytic residue identification *in vivo.*

The findings obtained from this thesis extensively expand the current scope of the DPP4 regulatory network, and this study is arguably the first attempt to systematically explore and characterise the DPP4 substrate repertoire in humans. DPP4-substrate structural modelling further improved our understanding of DPP4 enzymatic specificity and identified the most likely substrate entry pathway to be the side-opening (a more spacious and less hindered structure). The introduction of the hybrid deep neural network-enabled sequence mining model C5KmerCNNBiLSTM provided a combination of superior accuracy and exploration potential in predicting and validating DPP4 substrates. Furthermore, the top-down matrix-free tandem mass spectrometry has the potential to be further developed into a useful high-throughput tool to rapidly identify peptide cleavage events from *in vivo* samples. Future development of the validation workflow to globally verify all the predicted DPP4 substrates *in situ* will be welcomed.