Machine Learning Approaches in the Identification of Genetic Protein Markers Characteristic of Mature Onset Cataracts

Introduction

Cataracts, the number one cause of blindness worldwide, is characterized by the aggregation of unfolded and misfolded insoluble protein conformations that render eye lenses molecularly massive and opaque, resulting in progressively cloudy and impaired vision. Eye lenses are particularly susceptible to the aggregation of mutated proteins due to its development *in utero*, which means that mature fiber cells do not have the protein synthesis and degradation mechanisms used during fiber cell maturation needed to replace old or damaged lens structural proteins, and therefore the cells that make up the lens must be able to maintain and stabilize its structure for the organism's entire lifespan. This is why the crystallin proteins that compose 90% of the eye's structure must be soluble and stable enough to sustain their original structure, as both are essential for the light conductibility and long term maintenance of the lens.

In functional eye lens, α- and βγ-crystallin proteins are responsible for the regulation and maintenance of lens structure as well as the prevention of mutated protein conformation aggregation. α-Crystallin in particular is an ATP independent chaperone that binds to unfolded or misfolded proteins and prevents their conformation from aggregating. However, the α-crystallin population is finite, and as defective βγ-crystallin proteins gradually begin to saturate the lens, the more overwhelmed the α-crystallin becomes, which may subsequently even begin to contribute to aggregation.

This depletion of regulatory crystallin proteins is superficially linked to age-related cataract development, but analyzing the causes that would contribute to significant mutated protein aggregation is equally if not perhaps more important. Some recent studies have linked aggregation to heavy metals copper (Cu(II)-6Gly) as well as iron (Fe(III)-6Cit), noting that their redox active properties have been shown to promote the oxidation of o-aminophenols, which degrades the lens's natural UV filter and has demonstratively denatured the native protein structure of bovine lens proteins over time (Tweeddale et al). UV damage itself has been linked to aggregation; a study using guinea pig lenses linked the photooxidative effect of UV radiation to protein deletion in the lens nucleus as well as the accelerated depletion of soluble crystallin proteins (Giblin Fj, et al). It has been more recently suggested, however, that there may be a genetic link to cataract development separate from congenital cataracts found in children and young adults. The Framingham Offspring Eye Study Group found a found that familial cataract development could be clustered together using a generalized estimating equations (GEE2) logistic regression model, concluding that genetic or environmental factors could be to blame. Another study using the classical twin study model demonstrated that 48% of nuclear cataract development and 53% to 58% of cortical cataract development could be linked to heritable factors (Hammond, C. J. et al). The Beaver Dam Eye study was able to attribute 35% of nuclear cataract development and 75% of cortical cataract development to one gene (Klein, B. E. K. et

al). Thus, the purpose of this study is to do further research on the relatively novel idea of a linkage between genetics and late stage cataract development in an attempt to identify proteins that have mutations linked to cataract development in adults and perhaps utilize such information in future early diagnosis and even treatment, something that does not yet exist outside of special eyeglasses and cortical cataract surgery.

Methods

The early part of this study will primarily consist of a meta-analysis of papers that have found a linkages between mutated conformations of particular proteins present in the lens area and general mature onset cataract formation. At the moment of writing, 160 potential papers have been found. After meta-analysis is finished, a machine learning model will be written in to see if cataract development can be classified and predicted according to associated proteins, latter of which will have been gathered from the initial meta-analysis. A linear regression model is expected due to a speculatory correlation between certain protein types and cataract development, but other classification models will also be used for comparative purposes. Data will be acquired through external studies and existing databases; no data is planned to be developed through this study. Currently there are 4 available datasets primarily concerning mice cataract development, but more datasets are expected to be found throughout the study. Through these methods, we hope to establish a correlation between genetic proteins and mature onset cataract development.

References

Helen J. Tweeddale, Clare L. Hawkins, Joane F. Janmie, Roger J. W. Truscott & Michael J. Davies (2016) Cross-linking of lens crystallin proteins induced by tryptophan metabolites and metal ions: implications for cataract development, Free Radical Research, 50:10

Giblin FJ, et al. UVA light in vivo reaches the nucleus of the guinea pig lens and produces deleterious, oxidative effects. Exp Eye Res. 2002;75:445–458.

The Framingham Offspring Eye Study Group, Familial Aggregation of Lens Opacities: The Framingham Eye Study and the Framingham Offspring Eye Study, *American Journal of Epidemiology*, Volume 140, Issue 6, 15 September 1994, Pages 555–564,

Hammond, C. J., Duncan, D. D., Snieder, H., Lange, M. D., West, S. K., Spector, T. D., & Gilbert, C. E. (2001, March 01). The Heritability of Age-Related Cortical Cataract: The Twin Eye Study.

Barbara E. K. Klein, Ronald Klein, Kristine E. Lee, Emily L. Moore, Lorraine Danforth, Risk of Incident Age-related Eye Diseases in People with an Affected Sibling : The Beaver Dam Eye Study, *American Journal of Epidemiology*, Volume 154, Issue 3, 1 August 2001, Pages 207–211,