Probing Norwalk-like virus presence in shellfish, using artificial neural networks

G. Brion*, S. Lingeriddy*, T.R. Neelakantan*, M. Wang*, R. Girones, D. Lees***, A. Allard**** and A. Vantarakis*****

* Dept of Civil Engineering, University of Kentucky, Lexington, KY 40506, USA (E-mail: *gbrion@engr.uky.edu*)

** Biology School, University of Barcelona, Barcelona, Spain

*** Centre for Environment, Fisheries and Aquaculture Science, Weymouth, UK

**** Umeå University Hospital, Umeå, Sweden

***** School of Medicine, University of Patras, Patras, Greece

Abstract A database was examined using artificial neural network (ANN) models to investigate the efficacy of predicting PCR-identified Norwalk-like virus presence and absence in shellfish. The relative importance of variables in the model and the predictive power obtained by application of ANN modelling methods were compared with previously developed logistic regression models. In addition, two country-specific datasets were analysed separately with ANN models to determine if the relative importance of the input variables was similar for geographically diverse regions. The results of this analysis found that ANN models predicted Norwalk-like virus presence and absence in shellfish with equivalent, and better, precision than logistic regression models. For overall classification performance, ANN modelling had a rate of 93%, *vs* 75% for the logistic regression. ANN models were able to illuminate the site-specific relationships between indicators and pathogens.

Keywords Artificial neural networks; indicators; logistic regression; shellfish; virus

Water Science and Technology **Vol 50 No 1 pp 125–129**

Introduction

Health risks associated with consumption of virally contaminated shellfish are well documented, as is the need for a more reliable viral indicator system by the industry (Rippey, 1994; Lee and Morgan, 2003). More studies are needed to understand the underlying relationships between harvest area water quality, shellfish type, treatment processes and viral presence. Fortunately, a multi-country study in Europe has collected vital data in an effort to relate viral contamination of shellfish with these variables. Analysis of these results has been reported earlier by Formiga-Cruz *et al*. (2002, 2003), with the database consisting of records from geographically diverse areas collected over 18 months. The resultant data were provided to our team for further examination with new artificial neural networks (ANN) modelling tools.

Materials and methods

A multi-country study in Europe collected vital data in an effort to relate indicators to viral contamination of shellfish. As described in Formiga-Cruz *et al*. (2002, 2003), 20 sites from four countries were selected for shellfish harvesting and analysis of the meat for the presence of human enteric viruses [adenovirus (AV), enterovirus (EV), hepatitis A virus (HAV) and Norwalk-like viruses (NLV)] by nested PCR over an 18-month period. Other microbial indicators in the shellfish tissue [*E. coli*, somatic coliphages (SP), F-RNA phages (FP), *B. fragilis* phages (BP)] were quantified. Additional information (country, area, season, mollusc type, sea temperature and depuration status) was added to the microbial results, and the entire resultant dataset was modelled for virus prediction by logistic

regression (468 individual observations; Formiga-Cruz *et al*., 2003). The pertinent results from this study are the ability of logistic regression to predict positive and negative classifications for Norwalk-like viruses groups I and II (NLVI and NLVII), along with an overall correct classification from a logistic regression using the variables *E. coli*, SP, FP, BP, mollusc type and country. Another logistic regression model with the addition of temperature did not improve classification of NLVI- or NLVII-positive samples (Formiga-Cruz *et al*., 2003).

As a comparison, a feed-forward neural network model with back-propagation training using the software "Neurosort" developed by some of the authors (Brion *et al*., 2002) was applied to the same database for prediction comparison with similar input variables to the logistic regression done by Formiga-Cruz *et al*. (2003) (*E. coli*, SP, FP, BP, country, mollusc type). The ANN modelling differed from the prior logistic regression modelling approach in the fact that the microbial data were not log-transformed and additional input variables (month, area and depuration) were included with *E. coli*, SP, FP and BP for ANN training. An ANN model was trained to predict the positive and negative presence of NLV (genogroups I and/or II). Shellfish without NLVI and/or NLVII presence were classified as "0", while those with either genogroup NLVI and/or NLVII were classified as "1". The best architecture found to model the combined data from all four countries was nine input nodes: 15 hidden nodes: to one output node with learning rate of 0.04, momentum of 0.04 and 2,200 iterations for training.

To investigate the relative importance of the input variables for logistic regression, significance was calculated for the logistic regression input variables (Formiga-Cruz *et al*., 2003), while the relative strength effect (RSE) was calculated according to Kim *et al*. (2001) to determine the relative importance of the ANN models' inputs for predicting the outputs.

For the results presented in this paper, all of the data-points were used to train and validate both the logistic regression and the ANN models. This was done so that direct comparisons could be made on equal datasets. ANN classification rates were calculated for all observations (467), NLV positive and negative observations (69 and 395 respectively). Classification rates for overall, NLVI and NLVII for the logistic regression were obtained from Formiga-Cruz *et al*. (2003).

Results and discussion

All countries grouped

The best logistic regression classification of both positive and negative viral presence was found for Norwalk-like virus genogroups I and II (NLVI, NLVII) with about a 75% classification overall (positives and negatives) without temperature as an input variable, classifying negatives with slightly more accuracy than positives (Formiga-Cruz *et al*., 2003). The ANN model with similar input variables correctly classified NLV-positive samples 73.9% of the time. This was comparable to the positives correctly classified by the logistic regression model for NLVI (74.3%) and better than found for logistic regression positive classification of NLVII (67.4%). The ANN model was superior to the logistic regression for predicting the absence of NLV in shellfish (97.2% NLV *vs* 75.6% NLVI and 75.8% NLVII) and for the overall correct classification rate (93.4% NLV *vs* 75.5% NLVI and 75.1% NLVII).

Both modelling approaches found country to be a significant input variable (Table 1). However, the ANN model used area to provide more definition for the model. As is discussed below, this was to be expected, due to the difference in average indicator values and climatic effects between the countries participating in the original sampling of shellfish. Analysis of the ANN model found the additional input variable: month, to have the largest

absolute RSE – greater than that of country or area (–13.3 *vs* –6.7 and 0.9 respectively). Month is a temporal variable that can be related to water temperature, indicator and pathogen survival, and flooding events.

Inclusion of month improved the ANN classification prediction, whereas inclusion of temperature in the second, logistic regression model of Formiga-Cruz *et al*. (2003) decreased predictive power for the positive NLVI and NLVII classifications and increased power for the negative ones. However, the ANN model with temperature included as an input variable was superior to the logistic regression model for the prediction of overall classification (93.8% NLV *vs* 84.0 NLVI and 81.7% NLVII); superior for the positive NLV presence when compared to the temperature-inclusive logistic regression (73.9% NLV *vs* 60.9% NLVI and 63.0% NLVII); and superior for the negative NLV presence (97.5% NLV *vs* 85.9% NLVI and 83.5% NLVII).

Looking at the relative importance as indicated by significant dependence of the regression coefficients of the microbial indicators selected for use, the logistic regression for the grouped data found that coliphages, both FP and SP, had significant dependence for predicting NLVI or NLVII but not *E. coli* or BP. The absolute values of RSE from the ANN model for SP and FP were greater than for those of *E. coli* and BP (1.2, –0.4 *vs* –0.1, and 0.1 respectively) and agreed with the logistic regression model. The logistic regression and ANN models were also in agreement with the relative unimportance of BP and *E. coli* as input variables for the prediction of NLVs from the combined dataset; the inclusion of these variables did not impact the ability of the ANN to predict classification. The relationship between *E. coli* and the phage that uses it as a host did not cause multi-collinearity problems as it would have done with traditional linear regression modelling. It would seem, based upon the agreement between the models, that ANN models and RSE values could be used to optimise variable selection for a more classic logistic regression to ensure that inputs were correctly identified.

When ranking the geometric means from all sites within a single country, the data obtained from the UK shellfish had higher coliform and bacteriophage concentrations than all other countries, with the exception of BP in Sweden. However, this was influenced by sampling site coverage. Sites were selected to cover a range of all the faecal contamination levels found in the diverse countries. The UK sampling plan included both EU class C and prohibited sites in order to investigate virus occurrence in polluted sites restricted or prohibited for commercial exploitation. Spain ranked lowest in shellfish coliform and bacteriophage concentrations, with the exception of SP in Greece. This was to be expected, as the sampling sites selected for Greece and Spain were reported to have the highest incidence of sampling days where the potential *E. coli* host concentration was <230 MPN/100 g. Applying a Kruskal–Wallis One-Way ANOVA on ranks to log-transformed microbial data provided more information on the pattern of differences in the microbial concentrations between countries (Table 2).

Clearly, the average microbial concentrations at sites selected for study in the UK were quite different from those found in Spain and Greece, but shared more similarity with those from Sweden. The UK also used a different viral concentration method (but the same procedure for measuring the indicator organisms) than the other countries, and this methodological difference must be kept in mind when considering joining data sets. This was only

Table 1 Values for relative strength effects (RSE) for the ANN model for all countries

	Parameter								
	Country	**Month**	**Area**	**Mollusc**	**Depuration**	***E. coli***	**SP**	**FP**	**BP**
RSE $\frac{\partial y_k}{\partial x_i}$	–6.7	–13.3	0.9	0.9	————	–0.1	1.2	–0.4	0.1

Table 2 Differences in microbial indicator concentrations by country

	UK	Sweden	Greece	Spain
UK				
Sweden	N,Y,N,Y			
Greece	Y,Y,Y,N	Y,Y,Y,Y		
Spain	Y,Y,Y,Y	Y,N,Y,Y	N,Y,N,N	

one factor that may have "individualised" the underlying patterns between the response to be modelled and the input variables: a unique characteristic that could be accommodated well with ANN modelling.

Spain–Sweden ANN comparison

To investigate site-specific differences, further ANN modelling was applied to the Spain and Sweden datasets separately for the prediction of NLV in shellfish. These countries were selected because they demonstrated the strongest ANN training patterns (100% classification of positive and negative NLV presence) yet were different in average indicator concentrations, climate, and weather-related flooding. Sweden and Spain were shown to be different in microbial indicator concentrations for all but SP (Table 2). Sweden had experienced an extraordinary situation with respect to heavy rains and flooding in the western parts (Henroth *et al*., 2002). After selecting the best-trained ANN model for each country, the absolute values of the RSEs were compared (Table 3). The differences between the relative importance of the input variables supported the observations of the differences between the average concentrations. If the pattern between NLV presence and NLV indicators was site specific, then ANN investigations may have helped to decide when the databases from different regions should be joined.

In Sweden, the most important input variables were the area, month and FP. BP and FP inputs added strength to the ANN models' predictions of NLV in Swedish mussels. *E. coli* was of the least relative importance to the Sweden ANN model. As reported by Henroth *et al*. (2002), factors related to the flooding conditions (water flow, area, temperature and salinity) were the most highly valued as measured by logistic regression coefficients. This was in agreement with the RSE values from a 6:7:1 node architecture ANN model trained on month, area, *E. coli*, SP, FP and BP. Area and month had the highest RSEs, reflecting the impact that flooding had on this dataset. Of interest was that both the logistic and ANN models rated FP as the most important microbial indicator, with SP and BP phages of lesser importance and *E. coli* of least importance for the prediction of NLV. In fact, the values for *E. coli* were a factor of 10 less in magnitude for both logistic and ANN models as compared to the three types of bacteriophage.

In Spain, the order of importance for the input variables was different from Sweden. To predict NLV presence in various Spanish shellfish by ANN, the coliphages SP and FP were of equal importance to the model (RSE = 1.9) and of slightly greater importance than month (RSE = –1.4). The dependence upon month showed some seasonal influence in the

Table 3 Values for relative strength effects (RSE) for ANN models: Spain *vs* Sweden

	RSE $\frac{\partial y_k}{\partial x_i}$							
	Month	Area	Mollusc	Depuration	*E. coli*	SP	FP	BP
Spain ($\times 10^{-16}$)	–1.4	–0.4	0.1	————	0.9	1.9	1.9	0.8
Sweden ($\times^{10-1}$)	–8.7	10.9	NA	————	0.7	1.9	6.3	–2.8

incidence of NLV. In agreement with Sweden, FP was nearly equivalent with month, according to the RSE values. Area was of little importance for the Spain model (RSE = –0.4), in contrast to Sweden that had area and month as the greatest relative model inputs (RSE = 10.9 and –8.7 respectively). Thus, some of the patterns between input variables and NLV presence were similar between the two countries (e.g. the relative unimportance of *E. coli*), but there were differences as well.

The differences in the underlying patterns between indicators and NLV presence in shellfish for Spain and Sweden suggested that ANNs could be used to provide more information on if, and when, databases from differing regions could be combined. Much more study is needed to prove this proposed use of ANNs, but the findings of this study, and our experience with other studies, suggested that this use of ANNs should be more widely explored in concert with traditional statistical methods.

Acknowledgements

This research was supported by the Commission of the European Communities, Agriculture and Fisheries (FAIR) Project CT98-4039 and the USEPA (STAR) Project R830376.

References

Brion, G.M., Neelakantan, T.R. and Lingireddy, S. (2002). A neural network based classification scheme for sorting sources and ages of faecal contamination in water. *Wat. Res.*, **36**(15), 3765–3774.

Formiga-Cruz, M., Allard, A.K., Conden-Hansson, A.C., Henshilwood, K., Hernroth, B.E., Jofre, J., Lees, D.N., Lucena, F., Papapetropoulou, M., Rangdale, R.E., Tsibouxi, A., Vantarakis, A. and Girones, R. (2003). Evaluation of potential indicators of viral contamination in shellfish and their applicability to diverse geographical areas. *Appl. Environ. Microbiol.*, **69**, 1556–1563.

Formiga-Cruz, M., Tofino-Quesada, G., Bofill-Mas, S., Lees, D.N., Henshilwood, K., Allard, A.K., Conden-Hansson, A.C., Hernroth, B.E., Vantarakis, A., Tsibouxi, A., Papapetropoulou, M., Furones, M.D. and Girones, R. (2002). Distribution of human viral contamination in shellfish from different growing areas in Greece, Spain, Sweden and the United Kingdom. *Appl. Environ. Microbiol.*, **68**, 5990–5998.

Henroth, B.E., Conden-Hansson, A.C., Rehnstam-Holm, A.S., Girones, R. and Allard, A.K. (2002). Environmental factors influencing human viral pathogens and their potential indicators in the blue mussel, *Mytilus edulis*: the first Scandinavian report. *Appl. Environ. Microbiol.*, **68**, 4523–4533.

Kim, C.Y., Bae, G.J., Hong, S.W., Park, C.H., Moon, H.K. and Shin, H.S. (2001). Neural network based prediction of ground surface settlements due to tunneling. *Computers and Geotechnics*, **28**, 517–547.

Lee, R.J. and Morgan, O.C. (2003). Environmental factors influencing the microbiological contamination of commercially harvested shellfish. *Wat. Sci. Tech.*, **47**(3), 65–70.

Rippey, S.R. (1994). Infectious diseases associated with molluscan shellfish consumption. *Clin. Microbiol. Rev.*, **7**, 419–425.